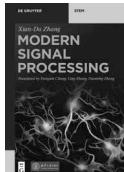


Ulrich Trick
5G

Also of interest



Modern Signal Processing

Xian-Da Zhang, 2022

ISBN 978-3-11-047555-5, e-ISBN (PDF) 978-3-11-047556-2



Blockchain

Technology and Applications for Industry 4.0, Smart Energy, and Smart Cities

Matevž Pustišek, Nataša Živić, Andrej Kos, 2022

ISBN 978-3-11-068112-3, e-ISBN (PDF) 978-3-11-068113-0

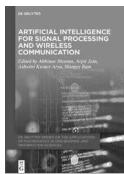


Metrology of Automated Tests

Static and Dynamic Characteristics

Viacheslav Karmalita, 2020

ISBN 978-3-11-066664-9, e-ISBN (PDF) 978-3-11-066667-0



Artificial Intelligence for Signal Processing and Wireless Communication

Hrsg. von Abhinav Sharma, Arpit Jain, Ashwini Kumar Arya, Mangay Ram, 2022

ISBN 978-3-11-073882-7, e-ISBN (PDF) 978-3-11-073465-2

in *De Gruyter Series on the Applications of Mathematics in Engineering and Information Sciences*

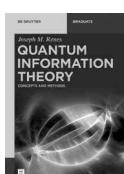
ISSN 2626-5427

Quantum Information Theory

Concepts and Methods

Joseph M. Renes, 2022

ISBN 978-3-11-057024-3, e-ISBN (PDF) 978-3-11-057025-0



Ulrich Trick

5G

The 5th Generation Mobile Networks

2nd Edition

DE GRUYTER
OLDENBOURG

Author

Prof. Dr.-Ing. Ulrich Trick
Frankfurt University of Applied Sciences
Research Group for Telecommunication Networks
Nibelungenplatz 1
60318 Frankfurt/M., Germany

ISBN 978-3-11-118648-1
e-ISBN (PDF) 978-3-11-118661-0
e-ISBN (EPUB) 978-3-11-118675-7

Library of Congress Control Number: 2023947507

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data are available on the Internet at <http://dnb.dnb.de>.

© 2024 Walter de Gruyter GmbH, Berlin/Boston
Cover image: koto_feja / iStock / Getty Images Plus
Druck und Bindung: CPI books GmbH, Leck

www.degruyter.com

In memoriam Cornelia

To the TDF+

Foreword

5G networks are being built and expanded worldwide. Compared to 4G, they not only offer obvious advantages such as higher bit rates but also high reliability and low latency, e.g., for car-to-x applications, or the integration of a very high number of end devices, such as sensors in a smart city.

To understand and sufficiently appreciate this, Chapters 1 and 2 follow the evolutionary development steps of mobile networks. This includes an overview of 2G and 3G with the different 3GPP (3rd Generation Partnership Project) releases, the introduction of the NGN (Next Generation Network) concept with VoIP (Voice over IP), the corresponding protocols SIP (Session Initiation Protocol), H.248 and Diameter, as well as the IMS (IP Multimedia Subsystem) to provide Multimedia over IP real-time services. A look at 4G with SAE (System Architecture Evolution) and LTE (Long Term Evolution), including VoLTE (Voice over LTE), completes the overview of the continuous development of mobile networks.

Starting with 4G, the increasing use of new network technologies such as NFV (Network Functions Virtualisation) and MEC (Multi-access Edge Computing), as well as SDN (Software Defined Networking) and SFC (Service Function Chaining), has become evident. Chapter 3 is dedicated to these essential technologies to implement the concept of so-called Future Networks and, consequently, 5G systems. The book's second edition provides enhancements and improvements on the topics of C-RAN (Cloud-Radio Access Network) and SDN.

The approach to 5G is different from previous mobile network versions, which were mainly driven by technology. As can be seen in Chapter 4, 5G started with possible use cases and deployment scenarios. Based on these, the requirements for the different application areas were derived, and only then the concepts and techniques required for the implementation were specified. The standardization is done in releases, as is usual with 3GPP. Releases 15, 16, and 17 have been completely standardized, Release 18 is in progress, and Release 19 has been started. In this context, as explained in Chapter 5, the ITU (International Telecommunication Union) should be mentioned in particular. It has defined a 5G target system based on the requirements and identified possible frequency ranges for 5G. These, in turn, have been and are still allocated to the network operators by regulators. The 2nd edition of the book brings updates on this, introducing, for example, Release 17, discussing the advantages and disadvantages of the various frequency ranges that can be used for 5G, extending the regulatory view to the EU and the USA, and providing an overview of the global 5G network rollout.

Chapter 6 provides an overview of a 5G system based on the applied design principles, the implementation features and associated functions, now also for Releases 16 and 17, and the resulting network architecture. Then, the technical details are worked out.

Chapter 7 provides deeper insights into the 5G access networks, focusing on the extremely powerful radio transmission technology, as well as discussing the topologies, architectures, and protocols of the RAN (Radio Access Network). A separate section on the increasingly important topic of O-RAN (Open-RAN) has been included in the 2nd edition of the book.

The highly innovative 5G core network is the subject of Chapter 8, where we discuss new topics such as Service Based Architecture (SBA) and Network Slicing. This central chapter has been significantly expanded in the 2nd edition of the book. It now contains descriptions of the signaling and user data protocol stacks. The usage of a UPF (User Plane Function) is discussed, especially for applications with high availability and latency requirements. In addition, new network functions introduced with Release 16 are explained, and the topic of network slicing is presented in more detail. In addition, the innovative opportunities for 3rd party providers through the network exposure functions with access to in-network functionalities are discussed. These extensions of Chapter 8 are completed by message captures from a real 5G system.

Chapter 9 summarizes the previously introduced concepts in an overall view, considering the 4G/5G migration, the use of the IMS in a 5G system, and the connection of various wired and wireless access networks up to satellite-supported base stations. The result is a network that implements FMC (Fixed Mobile Convergence) with only one core network technology. This is why 5G is not just a mobile network but also a new-generation converged network. The 2nd edition of the book complements this overall system view with two new sections on the topics of 5G and IoT (Internet of Things), including Time Sensitive Networking (TSN) and 5G campus networks. This chapter is concluded by an evaluation of 5G, including a comparison with 4G.

Since a 5G system is still an IP network, we must pay special attention to IT security by Chapter 10. A distinction is made between security for the communication network itself, security in the cloud infrastructure of the network operator, and the 3GPP security architecture standardized specifically for 5G. In the second edition, an update was made for the new releases and the Open-RAN.

This introduction to the 5th generation mobile networks is completed with an outline of the environmental influences due to electromagnetic radiation and the energy and raw material resources requirements in Chapter 11. Both subjects have been updated in the 2nd edition of the book. A new topic here is sustainability regarding a 5G system and its contribution to various areas of society and economic sectors. This is based on the 17 UN Sustainable Development Goals.

With the second edition, Chapter 12 finally provides a very detailed outlook on the future with the further development of 5G in Releases 18 and 19 at 3GPP, the work on a Network 2030 at the ITU, and activities and research results on 6G. In particular, 6G is examined in detail in four sections, starting with an overview of the organizations, initiatives, and research associations active in this area worldwide and regionally. The requirements for 6G are derived from use cases and deployment scenarios.

Subsequently, technologies and network architectures for implementation are discussed. In summary, a comparison of 6G with 5G is already being made at this early stage.

The book's main objective is to provide people interested in 5G technology and application scenarios with well-founded knowledge for an introduction to 5G and encourage further discussion of this topic. For this, the book refers to numerous additional sources in the 2nd edition, i.e., to 275 instead of 205 sources. The audience addressed by this book includes persons with a general interest in technology, primarily employees of public and private network operators. This book should be of particular interest within the IT departments of potential 5G user companies and, of course, among computer science and electrical engineering students. In addition, this book offers an optimally prepared introduction to the new topic of 6G.

For more information about this book, please visit the website www.5to6g.com. You are welcome to send me comments, suggestions, and questions by e-mail (trick@5to6g.com).

I especially want to thank my long-time companion, Prof. Dr. Armin Lehmann. On the one hand, he has contributed valuable technical impulses to this book, has been a constant discussion partner, and has critically reviewed the entire manuscript. On the other hand, together with M. Eng. Gregor Frick, he implemented a 5G system based on open source software and thus provided the already mentioned direct practical relevance. Many thanks to Gregor Frick for this support.

For the 2nd edition of the book, I would again like to thank Dr. Gerd Zimmermann, a proven 5G and experienced 3GPP standardization expert. Once again, he provided me with excellent support and advice.

Regarding the English edition, I would like to thank Gentiana Coman, Master in Computer Science. Without her comprehensive review covering all chapters, this English-language 5G book edition would not have been possible. I would also like to thank Dr. Besfort Shala, an expert in networks, IT security, and blockchain, for his helpful, careful, and critical review of the manuscript.

Last but not least, I would like to thank the De Gruyter publishing house. Dr. Damiano Sacco has initiated this 2nd edition of the 5G book; Mrs. Ute Skambraks was again very supportive during the complete book project; Mrs. Eva Kolla accompanied the work on this book up to the production. Many thanks to all of you for your support, excellent cooperation, and the well-designed cover.

Frankfurt/Main in October 2023

Ulrich Trick

Contents

1	Evolution of Mobile Networks — 1
1.1	Connection Concepts and Routing Principles — 4
1.2	Evolution of 2G/3G Mobile Networks — 10
1.3	NGN (Next Generation Network) — 15
1.4	VoIP (Voice over IP) and SIP (Session Initiation Protocol) — 19
2	3G/4G Mobile Networks and NGN (Next Generation Networks) — 30
2.1	3GPP Releases (3rd Generation Partnership Project) — 30
2.2	IMS (IP Multimedia Subsystem) and NGN — 32
2.3	H.248/Megaco Protocol — 40
2.4	Diameter Protocol — 46
2.5	SAE (System Architecture Evolution) and LTE (Long Term Evolution) — 58
2.6	VoLTE (Voice over LTE) — 61
3	Future Networks — 65
3.1	NFV (Network Functions Virtualization) and MEC (Multi-access Edge Computing) — 65
3.2	SDN (Software Defined Networking) and SFC (Service Function Chaining) — 74
3.3	Future Networks Concept — 95
4	5G Use Cases and Requirements — 99
4.1	5G Use Cases and Usage Scenarios — 99
4.2	Application Areas for 5G — 107
4.3	5G Requirements — 112
5	5G Standardization and Regulation — 121
5.1	Frequencies — 121
5.2	Standardization — 125
5.3	Regulation — 127
6	5G Networks at a Glance — 133
6.1	Design Principles — 133
6.2	Features and Functions — 135
6.3	5G Network Architecture — 142
7	5G Access Networks — 145
7.1	Radio Transmission Technology — 145

7.2	RAN (Radio Access Network) — 157
7.3	Open-RAN (O-RAN) — 163
8	5G Core Network — 167
8.1	Basic System Architecture and Protocols — 168
8.2	Core Network Functions — 178
8.3	Service Based Architecture (SBA) — 182
8.4	Network Slicing — 195
9	5G System — 204
9.1	4G/5G Migration — 206
9.2	5G and IMS — 209
9.3	Access Networks and Fixed Mobile Convergence (FMC) — 211
9.4	5G and IoT — 219
9.5	5G Campus Networks — 223
9.6	5G System in an Overall View — 226
10	5G and Security — 230
10.1	Security for the Communication Network — 234
10.2	Security in the Cloud Infrastructure — 236
10.3	3GPP Security Architecture for 5G — 242
11	5G and Environment — 249
11.1	New Issues through 5G Technology — 249
11.2	Electromagnetic Radiation and Health — 250
11.3	Exposure and Limit Values — 255
11.4	Influences of the Network Architecture — 257
11.5	Energy Requirements, Raw Materials, and Sustainability — 259
12	Future Developments — 267
12.1	Further Development of 5G — 267
12.2	Network 2030 — 271
12.3	Research, Regulation, and Standardization on 6G — 278
12.4	6G Use Cases and Usage Scenarios — 285
12.5	6G Requirements — 292
12.6	Technologies for 6G and Network Architectures — 296
Abbreviations — 307	
References — 323	
Index — 335	

1 Evolution of Mobile Networks

With 5G, the development of mobile networks has entered a new phase. In the past, the focus of such networks has been on the provision of communication services for people. In the case of 4G, multimedia data services such as video streaming with a smartphone, tablet, or generally a computer as the end device are the most important. With previous versions, the further back, the more the main focus was on telephony. Now, with 5G, the multimedia applications consumed by mobile users fall under traditional services, although supported by very high bit rates. Compared to previous versions, at least before 4G, the support of M2M (Machine to Machine communications) and IoT (Internet of Things) comes more into focus, but still with the corresponding 4G air interface, now with a high connection density compared to the beginnings with 4G. A completely new feature of 5G is the support of services in system and safety-critical application areas such as Smart Grid for intelligent energy supply networks and autonomous driving with very high demands on latency, response times, and system and service availability.

As shown in Figure 1.1, the introduction of digital mobile communications networks in the 1990s began with the 2nd generation – the 1st generation still used analog technology – based on GSM technology (Global System for Mobile Communications). Parallel to the GSM solution standardized in Europe by 3GPP (3rd Generation Partnership Project), the IS-54 (Interim Standard) and the IS-136, and finally, the IS-95 standard (cdmaOne) were developed in North America [271].

In many networks, the 2G solution consisted and still consists of a circuit-switched (CS) core network (CN) GSM and the associated access network (AN). With respect to the ease of use of IP over a mobile phone network, the CN was extended by a packet switching part, the GPRS (General Packet Radio Service). In parallel, the AN was migrated to be able to transport IP at moderate bit rates with EDGE technology (Enhanced Data Rates for GSM Evolution). This led to the current name GERAN (GSM/EDGE Radio Access Network).

In the early 2000s, the next step was the introduction of the 3rd generation, also known as UMTS (Universal Mobile Telecommunications System). Using W-CDMA (Wideband-Code Division Multiple Access) technology resulted in a much more powerful AN, UTRAN (Universal Terrestrial Radio Access Network), with significantly higher bit rates, but still with the CN based on GSM and GPRS. In the context of 3G, bit rates increased successively in the AN under the keyword HSPA (High Speed Packet Access).

There was also a parallel development in North America for 3G. The 3GPP partner organization 3GPP2 (3rd Generation Partnership Project 2) standardized the 3G cdma2000 solution with several successive versions [271].

The next step, the 4th generation, brought a new, high bit-rate access network technology based solely on IP, E-UTRAN (Evolved-UTRAN), under the name LTE

(Long Term Evolution). An LTE system provides telephony with VoIP (Voice over IP), here called VoLTE (Voice over LTE). Because of the real-time capability required for IP traffic, a new, real-time-capable IP core called EPC (Evolved Packet Core) became necessary. The IMS (IP Multimedia Subsystem), also shown in Figure 1.1 for the 3G evolution, is essential for signaling in VoLTE, and more generally, for Multimedia over IP services. The IMS with SIP (Session Initiation Protocol) plays an important role not only for 3G but also for 4G and 5G systems to provide real-time communication services.

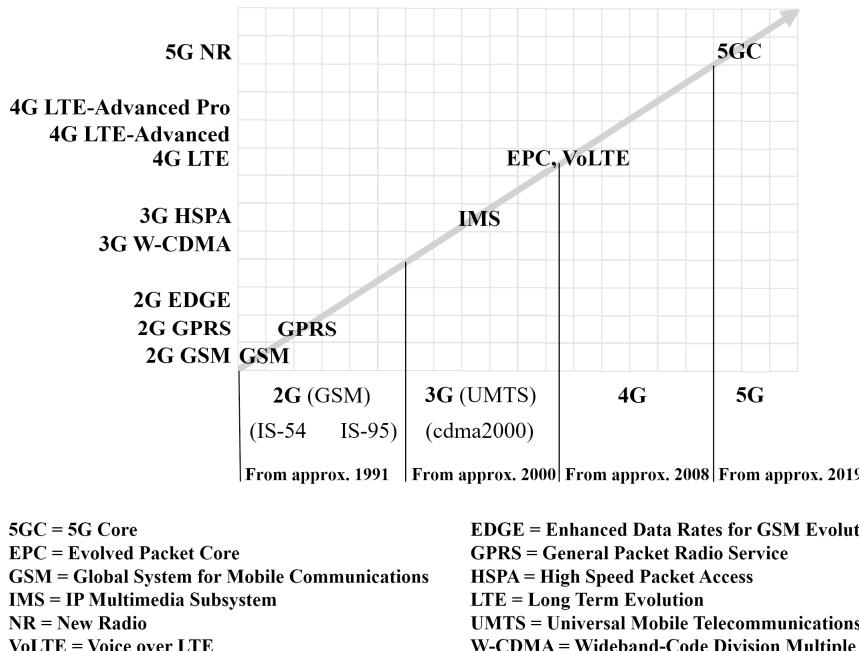


Fig. 1.1: Evolution of mobile networks [54]

The 4th generation of mobile networks is in operation today, alongside the parallel or integrated previous versions. It delivers high bit rates based on LTE, LTE-Advanced, and LTE-Advanced Pro access network technology and already has support for M2M and IoT with a separate Air Interface variant. In addition, the topic of virtualization with the use of only virtual network functions realized by software based on standard hardware has already started here [54].

The 5th generation of mobile networks is currently being launched and rolled out. It provides not only a new powerful RAN (Radio Access Network) technology, called NR (New Radio), for very high bit rates, very low delays (latency), and very high connection densities but also a new, highly modular, and flexible 5G core with

Service Based Architecture (SBA) and Network Slicing. The underlying technologies used are NFV (Network Functions Virtualization) and SDN (Software Defined Networking) in cloud environments. But this is not all. Without changing the core network, 5G also enables not only NR, non-3GPP WLAN, and 4G access but also fixed lines via, for example, PON (Passive Optical Network) or DSL (Digital Subscriber Line) and even direct access to a 5G network via a satellite connection. A 5G system can thus implement FMC (Fixed Mobile Convergence) with only one core network technology. For this reason, 5G can no longer be called a mobile network. If a 5G system is deployed and used in this general way, it is a new generation converged network.

The following sections and chapters deal with this evolution and to some extent revolutionary development. There is a good balance between introducing the basic ideas, concepts, and techniques, and more detailed considerations. We start with the basics, connection concepts, and routing principles. On this basis, the 2G/3G evolution is explained, and the NGN concept (Next Generation Networks), including VoIP and SIP, is covered. Chapter 2 describes concepts, protocols, and techniques of 3rd and 4th generation mobile networks. It includes IMS and VoLTE. Chapter 3 introduces the future networks standardized by the ITU. With NFV, Cloud, and Edge Computing, as well as SDN, they are already defining essential building blocks for 5G, anticipating 5G systems. From chapter 4 to chapter 10, there is a systematic introduction to 5G with more in-depth coverage wherever useful and necessary. The starting point is not new technical possibilities but use cases and new usage areas. It results in the requirements. These have been and still are the basis for standardization, especially in ITU and 3GPP, and regulation in individual countries. The requirements result in necessary network functions, which, according to selected design principles, lead to a 5G system and a 5G network architecture. For a more detailed analysis, a distinction can be made here between the access network and the core network. The knowledge gained in this process then leads to an overall view of a 5G system, including the interaction with 4G. Finally, concerning the technology, the security in a 5G system is considered.

Introducing a new network generation must also be considered from the perspective of the impact on the environment. Therefore, we address the topics of non-ionizing radiation due to radio transmission, energy consumption, and sustainability. Finally, we look into the future, first at the further development of 5G and then at an already planned 6th generation, which is currently in the research phase. That makes sense, as Figure 1.1 shows that a new mobile network generation is introduced approximately every ten years and that research, standardization, and development of the next network generation is already taking place parallel to the generation currently in operation.

1.1 Connection Concepts and Routing Principles

The technical development and, thus, the migration of the telecommunication networks and especially of the mobile networks, can be well characterized by the connection concepts and routing principles applied in each case.

As an introduction to this topic, Figure 1.2 shows an example of a connection setup for a telephone call between two subscribers (Sub) A and B in a telecommunications network or, more generally, in a Public Switched Telephone Network (PSTN). In addition to the architectural sketch of the network with the indicated switching centers (Exchange, Ex), a Message Sequence Chart (MSC) shows the time sequence of the signaling messages on the analog subscriber interfaces for establishing and terminating the connection [121; 161]. It is noticeable that there are three phases in the entire communication process for this telephone call:

- Connection establishment
- User data exchange
- Connection termination.

All three phases take place separately in time and sequentially. Therefore, this is called connection-oriented communication. However, this is only one possible and still imprecisely characterized connection concept.

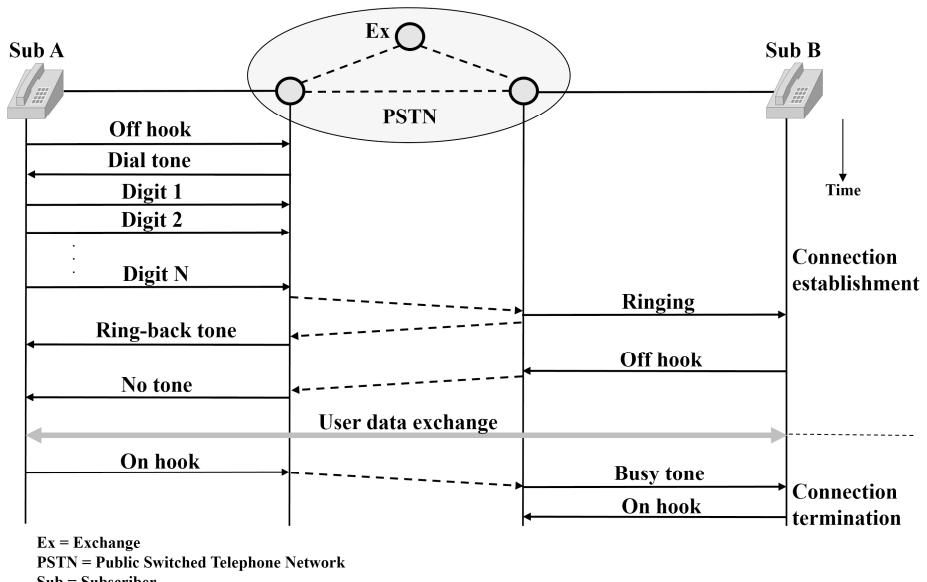


Fig. 1.2: Connection-oriented communication using the example of a telephone call in a PSTN

More generally and comprehensively, Figure 1.3 describes the term connection. At first, of course, the case of connection-oriented communication outlined above is shown here, with the connection phases represented by solid lines. In addition, this illustration also makes it clear that a connection concept can also include the partial or even complete overlapping of the three-time phases [121]. Based on this generally applicable description, the three connection concepts most relevant for characterizing telecommunications networks are worked out below.

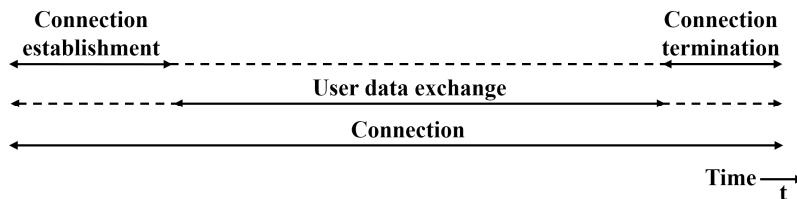


Fig. 1.3: The term connection in a generally applicable representation

As already mentioned, in connection-oriented communication, connection setup, user data exchange, and connection termination occur one after the other in terms of time. Another characteristic – well known from the practical example of the telephone call in Figure 1.2 – is the fact that during the connection setup, not only the target communication partner B is selected and informed about the possible connection, but they have the choice to accept or reject the connection [121; 161].

Taking the above explanations and the possible different forms of user data transmission into account, one can distinguish between three main connection concepts [121; 161].

Connection-oriented communication with a physically switched circuit

This first connection concept is connection-oriented. The user data (U) are transmitted in one or more physically switched channels provided by the network during the connection establishment phase (Ce). They are exclusively available to subscribers A and B until the connection is terminated (Ct). Figure 1.4 shows this connection concept. We use it, for example, in PSTN and ISDN (Integrated Services Digital Network) fixed networks or GSM-based 2nd and 3rd generation mobile networks. In all three cases, the user data exchange takes place via 64 kbit/s channels.

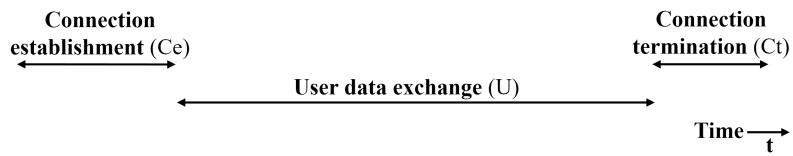


Fig. 1.4: Connection-oriented communication with a physically switched circuit

Connection-oriented communication with virtual circuit

The second connection concept also works connection-oriented. However, the channels for the exchange of user data are only occupied here if they are actually required. In this case, one speaks of virtual circuits. During periods of non-use (e.g., by A and B), other subscribers (e.g., C and D) can use the channel capacity or, more generally, the then free network resources. Figure 1.5 illustrates this, especially the greater flexibility compared to Figure 1.4. However, this means that the user data must be transmitted in the form of blocks (payload) with additional address and control information (header), not as continuous data streams. For this reason, the asynchronous time-division multiplex method shown in Figure 1.6 must always be used for this connection concept. The best-known application example for this second connection concept is an ATM (Asynchronous Transfer Mode) network with ATM cells of fixed length (5 Byte header and 48 Byte payload) for data transmission.

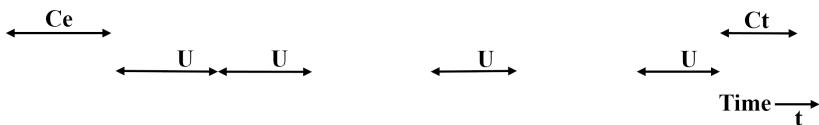


Fig. 1.5: Connection-oriented communication with virtual channel

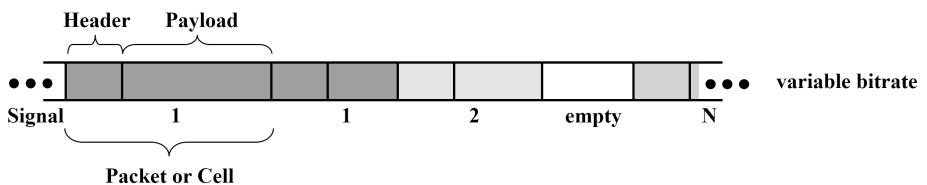


Fig. 1.6: Asynchronous time division multiplex

Connectionless communication

The third connection concept makes full use of the possible overlaps shown in Figure 1.3. According to Figure 1.7, the user data exchange (U), as well as the connection establishment (Ce) and disconnection (Ct), take place quasi simultaneously. A

user data block is transmitted from A to B without B being informed in advance and without the route from A to B being determined by the network. Connection establishment, user data exchange, and connection termination have overlapped in time each time a user data block has arrived at B. This also requires that the user data must be transmitted in the form of blocks (payload) with additional address and control information (header). We then speak of datagrams, which are, of course, also transmitted interleaved with the asynchronous time multiplex method, according to Figure 1.6. The best-known example of the use of this third connection concept is IP (Internet Protocol), with IP packets of variable length. Figure 1.8 shows their generation from a continuous user data stream.

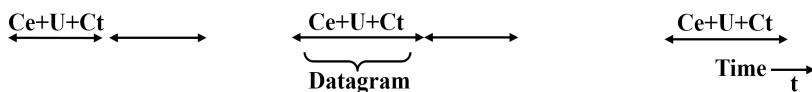


Fig. 1.7: Connectionless communication

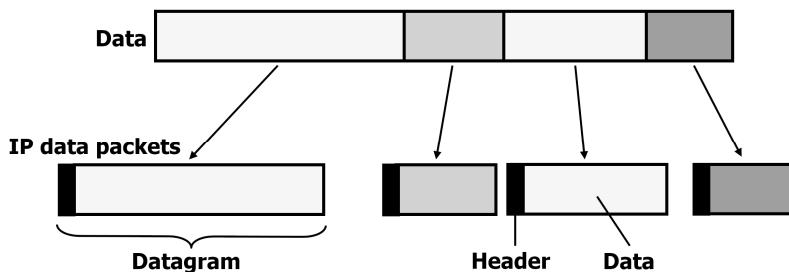


Fig. 1.8: Generation of datagrams or IP packets

Based on these three connection concepts, the routing principles can now be described [165].

Circuit Switching

Circuit switching is characterized by the fact that connection-oriented communication with physically switched user data channels is used. One of the consequences of this is that the bit rate for the user data is fixed (e.g., 64 kbit/s) for the duration of a connection, and the transit time through the network is constant and relatively low, which is a unique advantage for real-time services such as telephony. Figure 1.9 shows that the underlying multiplex method is the synchronous time division multiplex. The user data is transmitted with a fixed bit rate (e.g., 64 kbit/s), time-interleaved in time slots of fixed length (e.g., 8 bit = 1 Byte), and combined in frames (e.g., 32 time slots with 64 kbit/s each → 2,048 Mbit/s). Long-standing fields of ap-

plication for circuit switching are PSTN (see Figure 1.10), ISDN, and GSM. The solid lines in Figure 1.10 indicate the path for the 64 kbit/s user data that is physically switched during the connection setup.

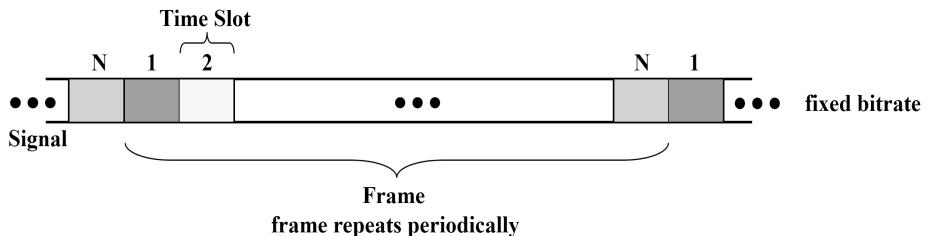


Fig. 1.9: Synchronous time division multiplex

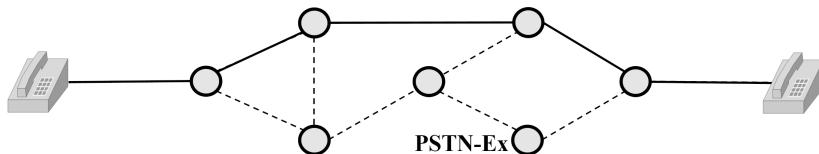


Fig. 1.10: Routing principle Circuit Switching using the example of PSTN

In so-called packet switching, as already mentioned above, the user data is transmitted in the network in the form of blocks with additional headers for address and control information. The cells (blocks of fixed length, e.g., ATM cells) or packets (blocks of variable length, e.g., IP packets) must be temporarily stored in the switching systems or routers to be evaluated and then forwarded. Therefore we also speak of store and forward switching. This is one of the reasons why runtimes vary. The bit rate can be adjusted according to the needs of the service. The transport capacities in the network can be flexibly allocated to different services or users and thus used in an optimized way. The asynchronous time division multiplex method shown in Figure 1.6 is used here [121; 161; 165].

In packet switching, a distinction is made between two variants according to the connection concept.

Virtual Circuit Packet Switching

Virtual Circuit Packet Switching works connection-oriented with virtual circuits. In the connection establishment phase, the network defines the path for the user data, but the data is only transmitted flexibly in the form of cells or packets if required. It also means that the user data always takes the same path through the network during a connection. The order of the cells or packets remains the same. This routing

principle is illustrated in Figure 1.11, using an ATM network. In addition to ATM applications, modern MPLS technology (Multiprotocol Label Switching) should also be mentioned here, in which MPLS frames of variable length transport IP packets. Of course, a connection in a network can also be established by configuration, not only by signaling [165].

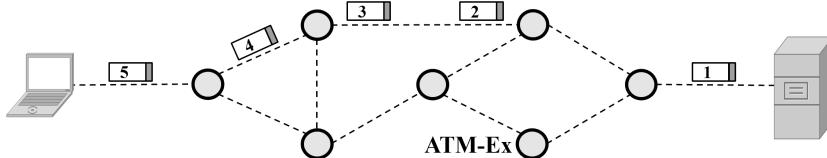


Fig. 1.11: Routing principle Virtual Circuit Packet Switching using the example of an ATM network

Datagram Packet Switching

This third routing principle works connectionless. It means that the path through the network is determined anew for each block or packet. This allows related packets, e.g., within a website request, to take different paths through the network. As a result, the sequence at the receiver may be different from that at the transmitter. An advantage of Datagram Packet Switching (mentioned in [121; 161] as Message Switching) is the high flexibility of the services with their different bit rate requirements and the optimized utilization of network resources. Besides, a network based on this routing principle provides optimum availability because as long as there is at least one possible path between source A and sink B, a packet is transmitted along this path. Datagram Packet Switching is used, as shown in Figure 1.12, in IP networks, including, of course, the Internet [165].

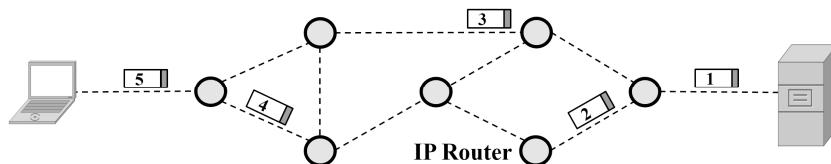


Fig. 1.12: Routing principle Datagram Packet Switching using the example of an IP network

1.2 Evolution of 2G/3G Mobile Networks

Using the routing principles described in Section 1.1, we can trace the development of 2nd and 3rd generation mobile networks.

2G GSM networks initially operated only with circuit switching for both voice and low bit-rate data services. In the second step, the GSM technology was extended by GPRS (General Packet Radio Service). It integrated an IP network based on Datagram Packet Switching to provide IP data services and interconnection to the Internet. The next step in network evolution was 3GPP Release 99 (3rd Generation Partnership Project) 3G-UMTS (Universal Mobile Telecommunications System), with circuit switching in GSM and datagram packet switching in the GPRS core network and an access network supporting higher bit rates. Operating two core networks with entirely different routing principles is not very efficient and, therefore, costly. For this reason, an IP transport network with media gateways for integration and transport of the 64 kbit/s user data channels of the GSM network, which is circuit-switched, was also specified for the GSM part of 3GPP Release 4. This step in the evolution of mobile networks has resulted in a combination of the routing principles Circuit Switching (GSM Core), Virtual Circuit Packet Switching (GSM Transport Network with controlled media gateways), and Datagram Packet Switching (GSM Transport Network and GPRS Core) with a tendency towards an ever-increasing share of Datagram Packet Switching and thus IP.

These evolutionary steps, which have only been briefly outlined so far, will be examined in more detail below for UMTS based on 3GPP Release 99 and 3GPP Release 4.

Figure 1.13 shows the network architecture, including some essential protocols for 3GPP Release 99, whereby for reasons of simplification, we consider only the GSM part of the UMTS core network here.

The exchanges of the circuit-switching GSM core network are the Mobile Switching Center (MSC) and the Gateway-MSC (GMSC) for the transition to other connection-oriented networks. The MSCs in Figure 1.13 correspond to ISDN exchanges with mobile-specific software. They communicate with each other in 64 kbit/s channels via a channel-oriented transport network. The transport of user data, e.g., for voice, still takes place in 64 kbit/s channels. The central signaling system No. 7 protocols ISUP (ISDN User Part) and TUP (Telephone User Part), supplemented by a mobile-specific part, are responsible for the exchange of messages to establish and terminate connections and to control services and supplementary services; MAP (Mobile Application Part) protocol is in charge of mobility control. These protocols are also used to connect to other mobile networks (using ISUP), the intelligent network (using the INAP (Intelligent Network Application Part) and CAP (CAMEL Application Part) protocols), and the ISDN or PSTN fixed network (using ISUP or TUP).

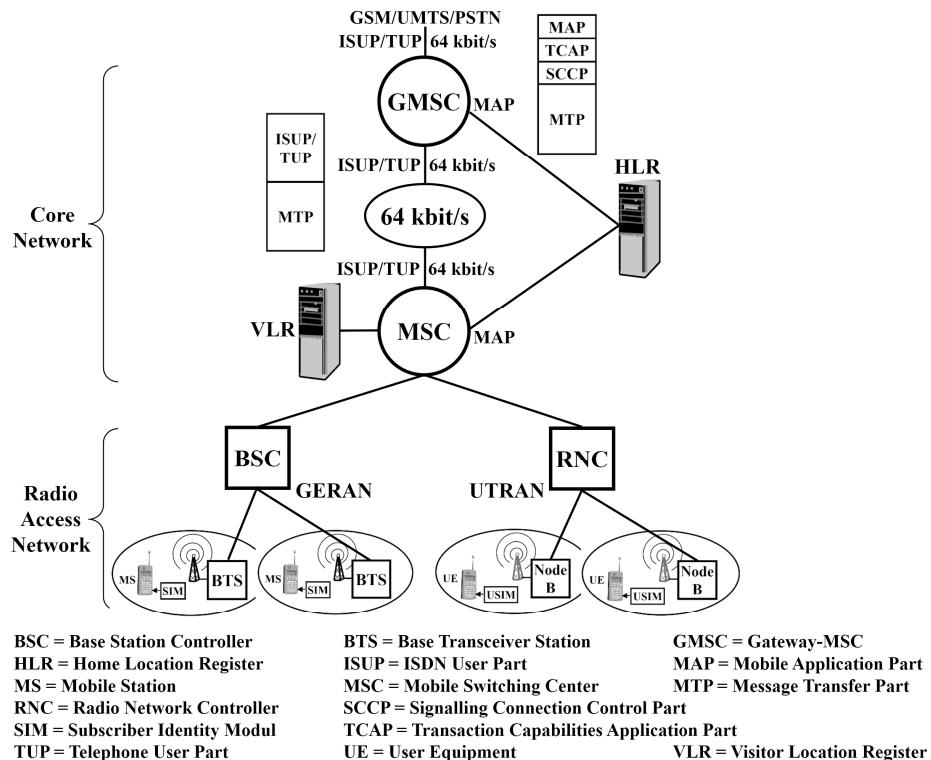


Fig. 1.13: 3GPP Release 99 UMTS mobile network with the focus on the circuit-switched GSM core network

To support comprehensive mobility within the network and also between GSM networks, the switching centers (the MSCs) can or must query various registers (i.e., databases) in the network: the Home Location Register (HLR), the Visitor Location Register (VLR), the Authentication Centre (AuC), and the Equipment Identification Register (EIR). The HLR contains the subscriber identification data, the services subscribed to by the user, the identification number of the MSC currently responsible for the subscriber, and, if necessary, the parameters for service features such as call forwarding. The VLR is generally linked to an MSC and contains a copy of the HLR data for all subscribers for which the MSC is currently responsible. The personal network access key is stored in the AuC for each subscriber. It is used for checking the network access authorization by authentication. The registration numbers of mobile stations (MS), e.g., smartphones, are managed in the EIR. It allows, e.g., identification as well as blocking of stolen end devices.

The 2G access network GERAN (GSM/EDGE Radio Access Network), which is also supported by a 3GPP Release 99 network, contains one Base Transceiver Station

(BTS) per radio cell. Several of these base stations are controlled by a Base Station Controller (BSC), a concentrator. Also, the BSC routes the corresponding traffic to the connected BTSSs.

The more powerful 3G access network UTRAN (Universal Terrestrial Radio Access Network), which in Release 99 supports bit rates of up to 2 Mbit/s per radio cell, is implemented using Base Stations Node B and the associated controllers RNC (Radio Network Controller).

Figure 1.13 shows the described UMTS network architecture, including the complete protocol stacks for ISUP/TUP and MAP [173; 31; 158].

According to Figure 1.14, the GSM architecture for packet data transmission via GPRS primarily consists of two logical network element types: SGSN (Serving GPRS Support Node) and GGSN (Gateway GPRS Support Node). These are packet switching centers (routers) that communicate with each other over an IP network, i.e., via Datagram Packet Switching. The SGSNs are responsible for encryption termination, PDP context handling (Packet Data Protocol), and IP routing, including mobility support. A GGSN is responsible for the IP address allocation to the mobile terminals, represents the anchor point for the PDP contexts when the responsible SGSN changes due to mobility, and acts as an IP router at the border to other packet-switching networks. Besides, for the mobility of GPRS subscribers, the HLR must be extended by GPRS-specific data or subscriber profiles, the so-called GPRS register (GR). Also, the currently associated MSC/VLR and SGSN continuously exchange information on the GPRS user's location. The BSC in the access network was originally developed to handle circuit-switched 16 kbit/s (voice) channels. It must be extended by the PCU function (Packet Control Unit) because of the packet switching required for IP. Figure 1.14 also shows the complete protocol stack used for IP transport in the GPRS core network for the UMTS network architecture described [173; 31; 158].

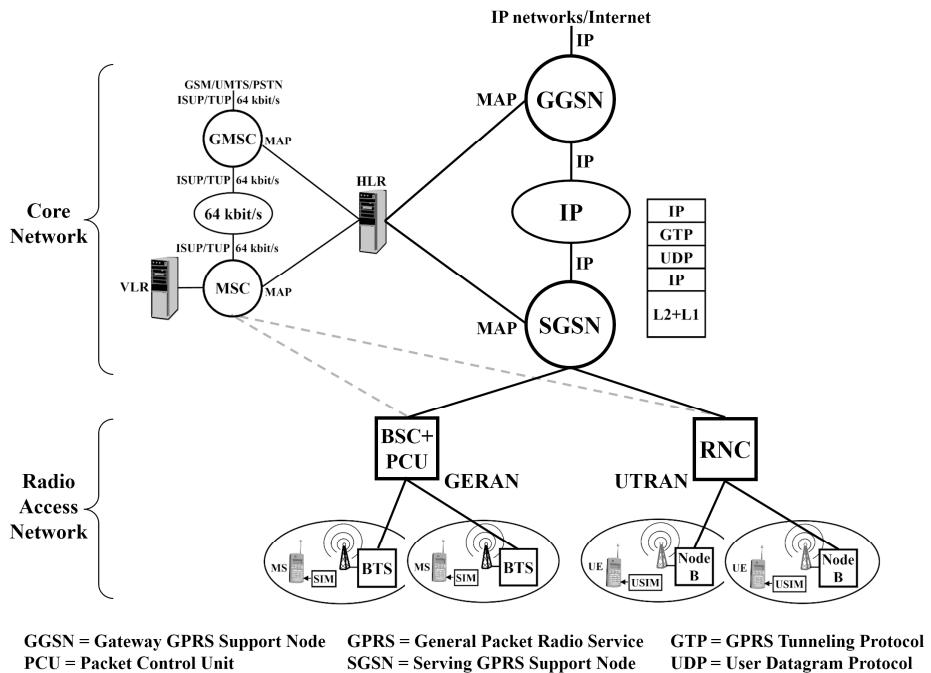


Fig. 1.14: 3GPP Release 99 UMTS mobile network with GPRS core network

As already mentioned above, the next evolutionary step with 3GPP Release 4 was the introduction of an IP transport network also for the GSM part based on circuit switching with 64 kbit/s user data channels. Figure 1.15 shows that the MSC and GMSC are divided into MSC/GMSC servers for signaling and network control, as well as circuit-switched Media Gateways (CS-MGW) for converting the 64 kbit/s real-time voice payload into VoIP RTP/IP packets (Voice over IP, Real-time Transport Protocol) and vice versa. MSC or GMSC servers, which are responsible for call control and mobility management, subsequently control the corresponding media gateways according to BICC (Bearer Independent Call Control) signaling using the H.248 protocol. The 64 kbit/s inputs and outputs of the media Gateways are defined according to the desired connections via H.248 messages. Here we can talk about Virtual Circuit Packet Switching. VoIP packets are exchanged between the MGWs. The IP network used for this is based on Datagram Packet Switching. Figure 1.15 shows the complete protocol stacks for BICC over SIGTRAN (SIGnalling TRANsport), H.248, RTP, and MAP, all based on IP [28; 158].

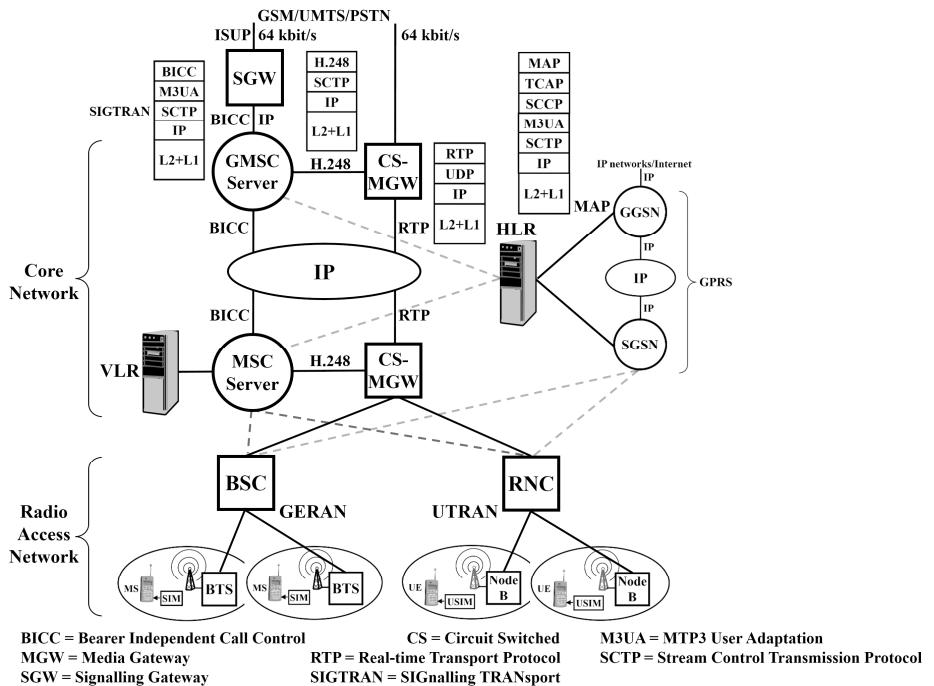


Fig. 1.15: 3GPP Release 4 UMTS mobile network

The network evolution of 2nd and 3rd generation mobile networks outlined above is leading to more and more IP:

1. GSM: Circuit Switching for voice and data
2. GSM + GPRS: Circuit Switching for voice and Datagram Packet Switching for data
3. 3GPP Release 99: Circuit Switching for voice and Datagram Packet Switching for data
4. 3GPP Release 4: Circuit Switching and Virtual Circuit Packet Switching for voice and Datagram Packet Switching for data.

The next and, therefore, 5th step was completed with 3GPP Release 5. This step introduced the NGN (Next Generation Networks) concept required for All over IP, initially for the provision of multimedia over IP services. Therefore, UMTS uses here: Circuit Switching and Virtual Circuit Packet Switching for voice and Datagram Packet Switching for data and multimedia.

The 3GPP Release 8 for a mobile radio network at the transition from the 3rd to the 4th generation finally leads to All over IP with a completely IP-based signaling and user data transport and thus Datagram Packet Switching for voice, data, and multimedia.

1.3 NGN (Next Generation Network)

The term NGN stands for a concept that can be described by the following points and the underlying network structure in Figure 1.16.

According to [173; 178], an NGN is characterized by:

- A packet-oriented (core) network for as many services as possible
- It includes real-time services such as telephony, so the network must provide a guaranteed Quality of Service (QoS).
- A particularly important point, both in terms of cost and openness to new services, is the complete separation of connection and service control from the transport of user data. The former is achieved with central Call Servers (CS). The main network intelligence is primarily implemented via software with cost-effective standard computer hardware. The latter offers the packet data network directly as well as gateways for the connection of channel-oriented operating networks, subnets, and end devices.
- By the NGN concept, all existing significant telecommunications networks, especially the technically different access networks which represent a high value, will be integrated. It is carried out with gateways for the user data (Media Gateway, MGW) and signaling (Signaling Gateway, SGW). Several MGWs are controlled by a central Call Server or the Media Gateway Controller (MGC) contained therein.
- For implementing value-added services, the Call Server communicates with application servers.
- Multimedia services and corresponding high bit rates are supported.
- Network integration aims not only at low system and operating costs through uniform technology, extensive reuse of existing infrastructure, optimal traffic utilization of the core network, and comprehensive uniform network management but also at general mobility.
- Integrated security functions ensure the protection of the transferred data and the network.

In addition

- an accounting system appropriate to the services,
- scalability,
- unrestricted user access to various networks and service providers, and
- the consideration of the applicable regulatory requirements (e.g., emergency call, lawful interception, security, privacy) must be ensured.

According to Figure 1.16, the gateway functionality can be part of the terminal device or the private circuit-switched network (residential gateway), represents the transition from the access network to the IP core network (access gateway), or con-

ncts a circuit-switched (e.g., ISDN) and a packet-switched (PS) core network (trunking gateway).

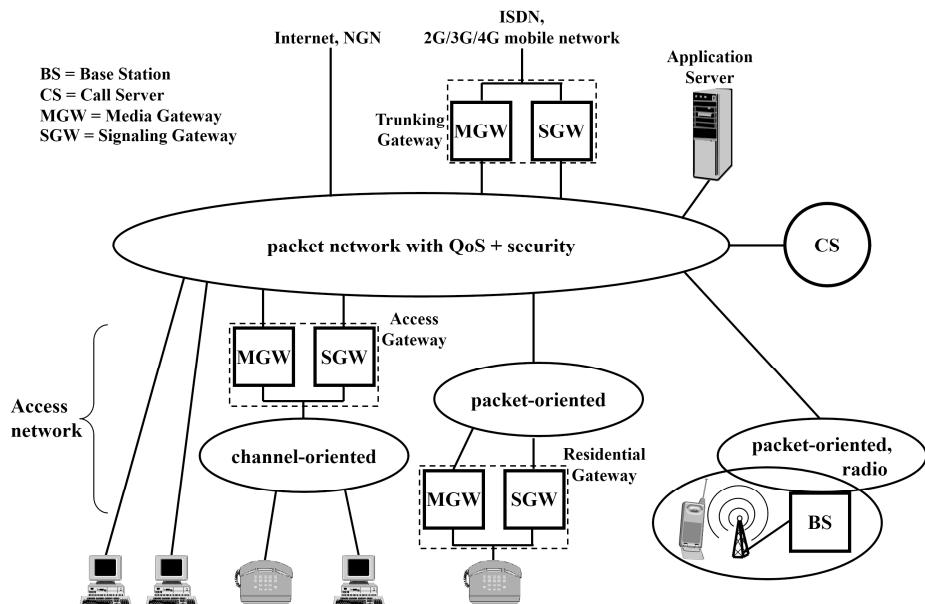


Fig. 1.16: Basic network structure of an NGN

The ITU-T (International Telecommunication Union-Telecommunication Standardization Sector) describes the concept “Next Generation Network (NGN)” in its definition in [178] briefly as follows: “A packet-based network able to provide telecommunication services and able to make use of multiple broadband, QoS-enabled transport technologies and in which service-related functions are independent from underlying transport-related technologies. It enables unfettered access for users to networks and to competing service providers and/or services of their choice. It supports generalized mobility which will allow the consistent and ubiquitous provision of services to users”.

The above compilation clarifies that the basic requirements for a modern telecommunications infrastructure are covered mainly by the NGN concept. The implementation of this concept seems to make sense for purely cost reasons if a network is to be newly implemented or extended or if it has to be modernized. At least in the core network, a network operator then manages only one IP data network instead of a separate network for voice and data. In addition, concerning the required bandwidth, data services are dominant anyway, and for them, the network is optimally adapted from the outset. Overall, this approach leads to fewer network elements,

more uniform technology, unification of network management, and, thus, to cost savings in procurement and, above all, in operation. In addition, new services, especially multimedia services, can be implemented more easily with the integration of voice and data than in legacy networks [173].

The outlined concept does not specify the applied protocols. However, today packet networks are always implemented based on IP, so the connectionless IP has been established for an NGN. In addition, however, a signaling protocol for call control is required for services such as telephony. For this purpose, the SIP (Session Initiation Protocol) has established itself worldwide, supported, among other things, by definition as a standard for 3GPP Release 5. Cooperating protocols like SDP (Session Description Protocol), RTP (Real-time Transport Protocol), and H.248/Megaco complement SIP and IP [173].

Figure 1.17 shows the principal structure of such an IP-based network, in which the connection and service control is implemented using SIP. If a SIP User Agent (e.g., a PC that works as a softphone with appropriate telephone software) wants to connect to a telephone (in this case, an IP phone) via the IP network, it uses SIP to establish the desired connection (after registering with a SIP Registrar server) via a SIP Proxy server and other proxy servers if necessary. The media parameters for the user data are negotiated via SDP. After setting up the SIP session, RTP (Real-time Transport Protocol) sessions are established for the packaged voice user data.

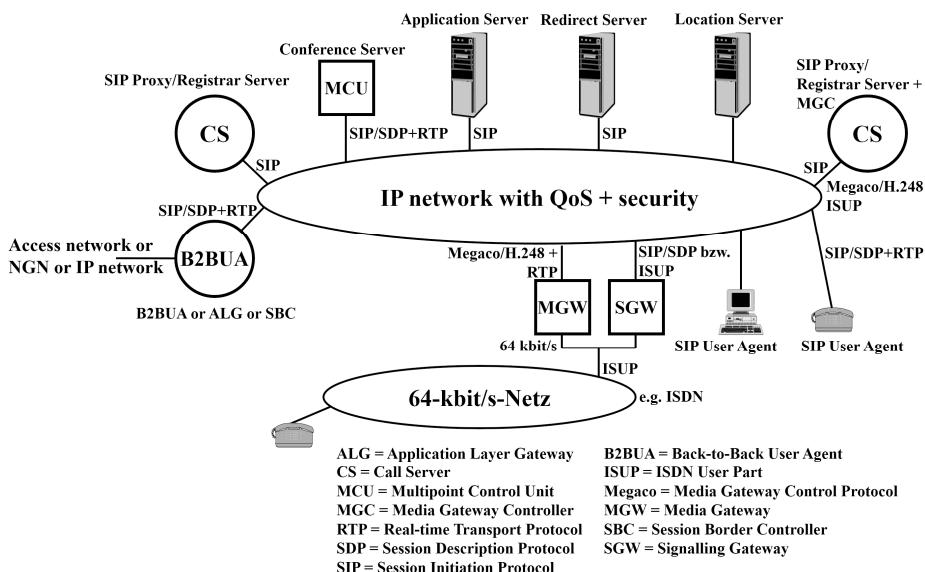


Fig. 1.17: Protocols and network architecture for NGN with SIP for signaling

The Location server stores the relationship between the permanent and the IP subnet dependent temporary SIP addresses. It receives this information from the Registrar server and makes it available to the SIP Proxy server for session control. A Redirect server also provides mobility support by providing a calling SIP User Agent with alternative destination addresses of the called party.

The communication, e.g., into ISDN, is realized via gateways, whereby the actual gateway (MGW + SGW) and the control of the MGW are separated according to the NGN concept. The controller, the Media Gateway Controller (MGC), is part of the Call Server functionality. It communicates with the MGW via protocol H.248/Megaco. The application servers are used to implement value-added services. They work together with the SIP proxy servers via SIP. The Conference server/MCU (Multipoint Control Unit) supports conferences.

The different server types (e.g., SIP Registrar server, SIP Proxy server, Media Gateway Controller) represent logical units. Physically, they can be implemented stand-alone or in combination.

As described above, the gateway elements MGW and SGW work in close cooperation with the Call Server and the Media Gateway Controller, respectively. The Media Gateway (MGW) only realizes the conversion between 64 kbit/s user data channels and IP packets; it is entirely remote-controlled by the MGC via Megaco/H.248. Both standards – H.248 and Megaco – describe the same protocol. The Signaling Gateway (SGW) typically only converts the protocols for the transport of signaling messages, not the signaling itself. In the case of connecting a digital telephone network with ISUP signaling to an IP network with SIP signaling, the SGW only converts the lower protocol layers MTP (Message Transfer Part) to IP in combination with SCTP (Stream Control Transmission Protocol). The ISUP messages are transmitted transparently to the Call Server, and only there is a conversion to SIP. It is the typical gateway application in public and, thus, larger networks. In these cases, one speaks of decomposed gateways. The user data conversion takes place in the MGW, that of the signaling messages in the MGC, i.e., in separate devices. The situation may be different in private and, therefore, often small networks. Here, MGW and SGW are usually combined in one device and appear in the direction of ISDN as ISDN terminal and in the direction of the IP network as SIP User Agent.

At the transition between two SIP/IP networks, e.g., between the core and access networks of a provider or between two NGNs of different network operators, there is often a requirement to analyze and, if necessary, process both the SIP signaling and the RTP user data. Reasons for this can be security requirements, necessary IP address translations, hiding the network topology, providing anonymity, etc. The SIP network elements B2BUA (Back-to-Back User Agent), ALG (Application Layer Gateway), and SBC (Session Border Controller), also shown in Figure 1.17, serve this purpose. In practice, the first two are primarily applied in combination with other network elements as logical SIP network elements; SBCs are often used as stand-alone devices. All three mentioned network element types offer the same or

at least similar functions; only the focus is slightly different: mainly B2BUA for signaling handling, ALG for security functions and address translation, SBC at network-network interfaces [173].

1.4 VoIP (Voice over IP) and SIP (Session Initiation Protocol)

Based on the explanations in Section 1.3, Figure 1.18 summarizes the essential protocol stacks for a SIP/IP-based NGN. Based on this, the protocols SIP, SDP, and RTP, which are important for all 3GPP releases from 5 upwards, will be explained.

In modern IP networks, SIP is the protocol for signaling in session-based, i.e., connection-oriented communication for multimedia services. It also includes VoIP, which is crucial for telephony. SIP thus fulfills functions in IP networks such as ISUP or DSS1 (Digital Subscriber Signaling system no. 1) in ISDN. It also offers additional features such as the transmission of short text messages and status event monitoring, e.g., the presence status of a subscriber.

SIP has been standardized in numerous RFCs (Request for Comments) by the IETF (Internet Engineering Task Force). Especially relevant is the basic standard RFC 3261 [3]. Since it is an Internet protocol, functions were taken over from HTTP (Hypertext Transfer Protocol); the SIP messages are, therefore, purely text-based.

User data		Signaling					
Codec	RTCP	SIP (Register)	SIP (Call Signalling)	SDP (Bearer Control)			
RTP							
UDP		UDP (TCP)					
IP							
Layer 2 protocol							
Layer 1 protocol							

RTCP = RTP Control Protocol

Fig. 1.18: Protocol stacks for Multimedia over IP

As shown in the protocol stack in Figure 1.18, SIP messages are usually transported via the connectionless UDP (User Datagram Protocol). However, the connection-oriented TCP (Transport Control Protocol) or other Layer 4 protocols can also be used, depending on requirements.

SIP distinguishes between two address types, so-called URIs (Uniform Resource Identifier): first, a permanent SIP URI in the form `sip:user@domain` (e.g., `sip:trick@providerx.com`), which is permanently assigned to the user himself and can be compared with a telephone number. The domain identifies the SIP service provider, the user the individual subscriber. Secondly, there is another SIP URI in the form `sip:user@IP address:Port number` (e.g., `sip:trick@98.60.105.14:10503`), which temporarily identifies the end device applied by the user, the SIP User Agent (SIP UA), and addresses it in the current IP subnet and makes it accessible. The relationship between permanent and temporary SIP URIs, crucial for SIP routing, is determined in the SIP registration process, as indicated in Figure 1.18. As briefly sketched in Section 1.3, the SIP UA registers with the SIP Registrar server after its activation. The Registrar server captures the relationship between permanent and temporary SIP URI and stores it in the Location server. As a result, the SIP Proxy server has access to this information for routing operations [173].

There are two types of SIP messages: requests and responses (status information). A SIP request is specified by an English identifier, the so-called method, which gives a clear indication of the meaning of the protocol message. Table 1.1 shows the most relevant SIP request messages for understanding SIP [173; 3].

Tab. 1.1: Selection of important SIP request messages

SIP request	Function
INVITE	Initiating a SIP session (connection setup)
ACK	Confirmation of receipt of a final SIP response message as a result of an INVITE request
REGISTER	Registration of a SIP User Agent
BYE	Termination of an existing SIP session (connection termination)
MESSAGE	For short text messages
SUBSCRIBE	Initiation of event monitoring, e.g., to query the presence status of a user
NOTIFY	Feedback on a requested event, e.g., in case of changes in the presence status
PRACK	Provisional ACK, to interrupt a running transaction, e.g., to reserve resources for defined QoS

The so-called status code, a three-digit decimal number, identifies a SIP response. It is supplemented by a standard reason phrase, which, in contrast to the status code, is not binding and can, therefore, be changed on a case-specific basis. The number

of SIP responses is quite large, so they are divided into six basic types according to their primary functions. Table 1.2 provides an overview of these basic types and the most relevant SIP response messages from an understanding point of view. Here the connection with HTTP becomes clear again [173; 3].

Based on some of the above SIP requests and SIP response messages, Figure 1.19 shows a simple MSC for a peer-to-peer SIP session setup and termination. It becomes clear that the INVITE request is followed by three responses, of which the two 1xx messages are optional. The successful session establishment is signaled by 200 OK, which in turn results in the ACK request. ACK is the only request message that does not expect responses. Besides, it must always be sent as a confirmation after a previous INVITE and resulting responses 2xx and higher. This sequence, INVITE – 2xx, 3xx, 4xx, 5xx, or 6xx – ACK, is called SIP Three-Way Handshake. Session termination is initiated with the BYE request and also confirmed as successful with 200 OK [173; 3].

Tab. 1.2: Basic SIP response types and selection of essential response messages

Basic type	SIP Response
1xx (Provisional Responses)	100 Trying 180 Ringing 181 Call is Being Forwarded 183 Session Progress
2xx (Successful)	200 OK
3xx (Redirection)	301 Moved Permanently 302 Moved Temporarily
4xx (Request Failure)	401 Unauthorized 404 Not Found 407 Proxy Authentication Required 415 Unsupported Media Type 486 Busy Here
5xx (Server Failure)	500 Server Internal Error 503 Service Unavailable 504 Server Timeout
6xx (Global Failure)	600 Busy Everywhere 603 Decline

SIP request and SIP response messages have the same structure. As shown in Figure 1.20 as an example of an INVITE request, each SIP message consists of a start line with the method and the Request URI or status code with a reason phrase and a SIP message header with numerous header fields. If required, this is followed by an optional message body with, for example, an SDP message or a short text message.

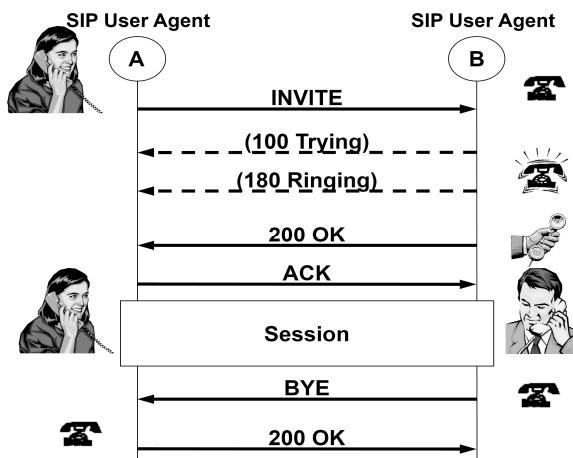


Fig. 1.19: Typical message exchange during SIP session setup and termination

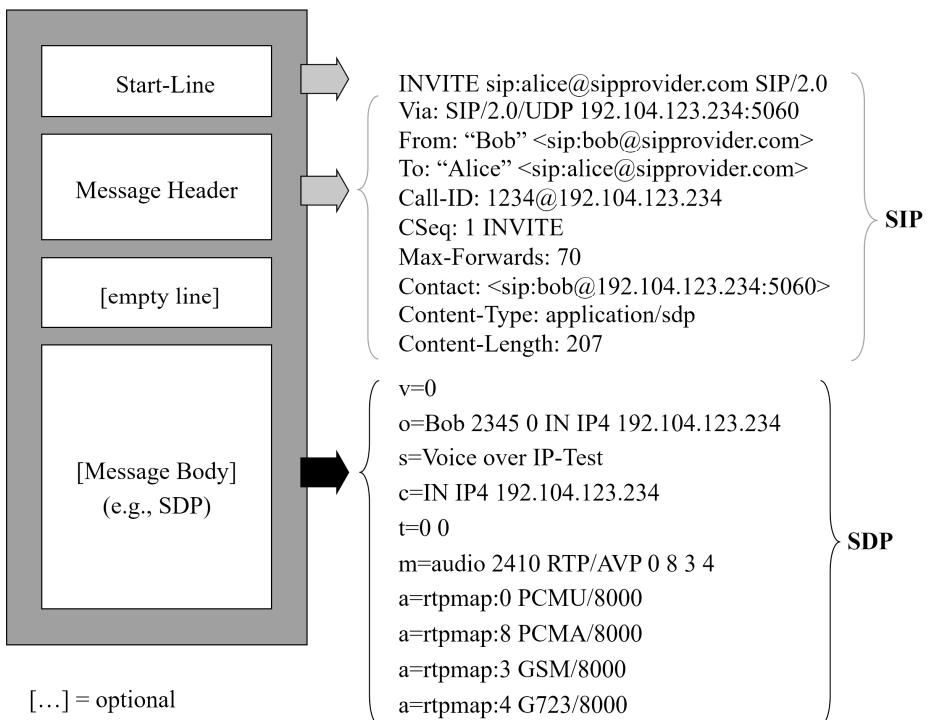


Fig. 1.20: Structure of SIP messages with message header and message body

Figure 1.20 shows relevant header fields for the exemplary INVITE message to the Request URI `sip:alice@sipprovider.com`, whose functions Table 1.3 explains. Header fields that go beyond the simple example in Figure 1.20 are also listed and explained. They are part of the two practical message records. Figure 1.21 presents an INVITE request to User Agent B (`bob`), which has already passed through two SIP proxy servers (3 x `Via`, 2 x `Record-Route`). Figure 1.22 shows an ACK request (2 x `Route`) sent by User Agent A (`alice`) to complete the SIP three-way handshake [173; 3; 117].

Tab. 1.3: Important SIP header fields and their functions

SIP Header field	Function
<code>Via</code>	Information for routing the SIP responses, including the socket (IP address:port number) of the request sender. Ensures that a response takes the same path as the initiating request
<code>From</code>	Permanent SIP-URI of the request sender, the transaction initiator, the so-called User Agent Client (UAC)
<code>To</code>	Permanent SIP-URI of the request receiver. Replies with response(s), so-called User Agent Server (UAS)
<code>Call-ID</code>	Identifies all SIP messages that belong to a session or dialog
<code>CSeq</code>	Identifies all SIP messages that belong to a transaction (request + all resulting responses)
<code>Max-Forwards</code>	Number of still allowed SIP hops to avoid endless loops
<code>Contact</code>	Temporary SIP-URI of the SIP User Agent sending the message
<code>Content-Type</code>	Data type in the message body, here SDP
<code>Content-Length</code>	Data length in Byte in the message body
<code>Expires</code>	Validity period in seconds of a SIP event, such as a registration
<code>Allow</code>	Methods supported by the SIP network element
<code>User-Agent</code>	Description of the User Agent
<code>Record-Route</code>	It means that a SIP Proxy server specifies for the first routed request that all subsequent requests of a session, including ACK, BYE, etc., must be routed through it
<code>Route</code>	Based on the received Record-Route headers for a new request, identifies all SIP proxy servers that must be traversed

```

Frame 11: 1185 bytes on wire (9480 bits), 1185 bytes captured (9480 bits)
Ethernet II, Src: CadmusCo_49:37:79 (08:00:27:49:37:79), Dst: CadmusCo_ba:e7:17 (08:00:27:ba:e7:17)
Internet Protocol Version 4, Src: 172.20.0.57 (172.20.0.57), Dst: 172.20.0.50 (172.20.0.50)
User Datagram Protocol, Src Port: 5060 (5060), Dst Port: 5060 (5060)
Session Initiation Protocol (INVITE)
Request-Line: INVITE sip:bob@172.20.0.50 SIP/2.0
Message Header
Record-Route: <sip:172.20.0.57;lr=on>
Record-Route: <sip:192.168.5.1;lr=on>
Via: SIP/2.0/UDP 172.20.0.57;branch=z9hG4bk83e6.1266784808935b5383a741d587016767.0
Via: SIP/2.0/UDP 192.168.5.1;branch=z9hG4bk83e6.eed8aacf09047792ac63158ad1552133.0
Via: SIP/2.0/UDP 192.168.5.17:5060;rport=5060;branch=z9hG4bk408046846
From: <sip:alice@192.168.5.1>;tag=1995507312
To: <sip:bob@172.20.0.57>
Call-ID: 117445833
CSeq: 20 INVITE
Contact: <sip:alice@192.168.5.17>
Content-Type: application/sdp
Allow: INVITE, ACK, CANCEL, OPTIONS, BYE, REFER, NOTIFY, MESSAGE, SUBSCRIBE, INFO
Max-Forwards: 68
User-Agent: Linphone/3.3.2 (exosip2/3.3.0)
Subject: Phone call
Content-Length: 405
P-hint: outbound
Message Body

```

Fig. 1.21: SIP INVITE request captured with protocol analysis software

```

Frame 25: 447 bytes on wire (3576 bits), 447 bytes captured (3576 bits)
Linux cooked capture
Internet Protocol Version 4, Src: 192.168.5.17 (192.168.5.17), Dst: 192.168.5.1 (192.168.5.1)
User Datagram Protocol, Src Port: 5060 (5060), Dst Port: 5060 (5060)
Session Initiation Protocol (ACK)
Request-Line: ACK sip:bob@172.20.0.50 SIP/2.0
Message Header
Via: SIP/2.0/UDP 192.168.5.17:5060;rport;branch=z9hG4bk864398346
Route: <sip:192.168.5.1;lr=on>
Route: <sip:172.20.0.57;lr=on>
From: <sip:alice@192.168.5.1>;tag=1995507312
To: <sip:bob@172.20.0.57>;tag=1358134314
Call-ID: 117445833
CSeq: 20 ACK
Contact: <sip:alice@192.168.5.17>
Max-Forwards: 70
User-Agent: Linphone/3.3.2 (exosip2/3.3.0)
Content-Length: 0

```

Fig. 1.22: SIP ACK request captured with protocol analysis software

Based on the above explanations of SIP requests and responses and the corresponding SIP header fields, we can now consider a complete SIP routing process shown in Figure 1.23. First of all, the two SIP User Agents A and B have to be registered with their permanent and temporary SIP URIs through a REGISTER message for a certain period of validity (Expires header field). Therefore this information is available on the Location server.

User Agent A creates an INVITE request to the permanent SIP URI of subscriber B (here: `sip:B@Provider.com`) specifying the Via header field for receiving SIP status information (here: IP address 87.87.87.87) and the temporary SIP URI in the Contact header field (here: `sip:A@87.87.87.87`) and sends it in step (1) to a SIP Proxy server (here: IP address 89.89.89.89). It queries the Location server for the temporary SIP URI (here: `sip:B@90.90.90.90`) registered under the permanent SIP URI of subscrib-

er B. Since the Proxy server wants to be included in every further step of the SIP signaling of this session, it automatically sets the Record-Route header field by specifying the IP address (here: 89.89.89.89) or a domain name resolvable by DNS. The Proxy server then adds a Via header field above the existing Via header field. It provides its contact parameters (here: IP address 89.89.89.89) for the back-routing of response messages answering the request.

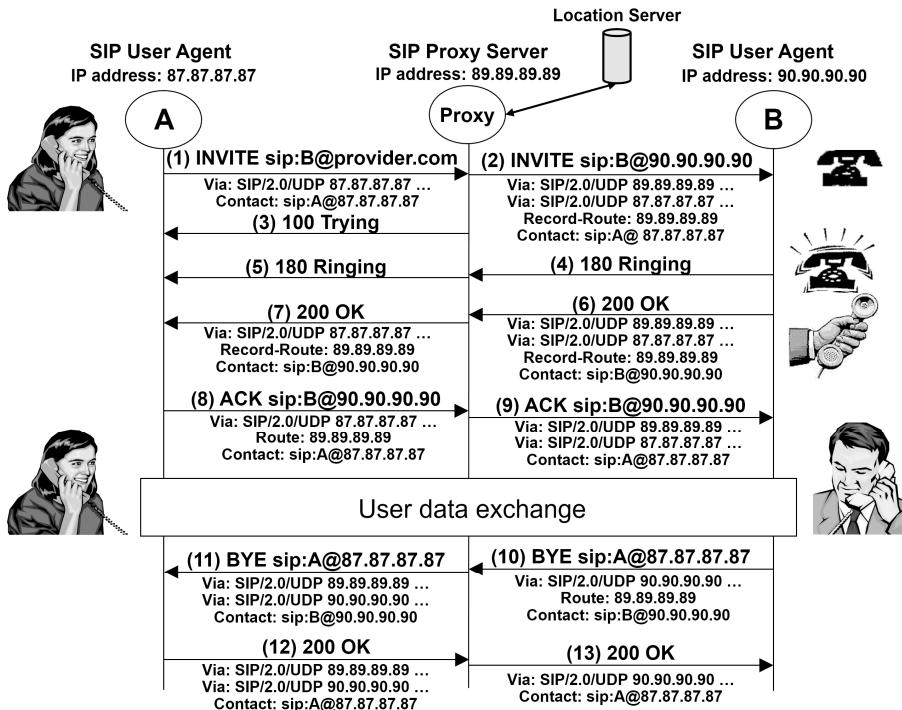


Fig. 1.23: Routing of SIP messages

In step (2), the Proxy server routes the INVITE request to User Agent B. After receiving the INVITE request, User Agent B sends the status information 180 Ringing (4) and 200 OK (6) to the contact address of the Proxy server specified in the uppermost Via header field (here: IP address 89.89.89.89). The Via and Record-Route header fields from the INVITE request are copied by User Agent B in the same order into each sent status information. Moreover, User Agent B has transferred his temporary SIP URI (here: sip:B@90.90.90.90) in the Contact header field of this response. Figure 1.23 shows the SIP header fields Via, Record-Route, and Contact as examples for the final status information 200 OK (see (6) and (7)) only. The Proxy server receiving the status information sent by B deletes the Via header field that identifies it from

the responses. The Proxy server leaves the Record-Route header field in the status information, which it forwards to the contact address specified in the remaining Via header field (here: IP address 87.87.87.87), i.e., to User Agent A (see (5) and (7)). After receiving the 200 OK response, signaling the session acceptance by subscriber B, the SIP transaction initiated with the INVITE request (1) is completed for User Agent A. To complete the SIP Three-Way Handshake, it must send the SIP message ACK to User Agent B.

In the first forwarded request INVITE (2), the Proxy server previously involved in the SIP signaling exchange has used the Record-Route header field to announce that it will remain in the SIP signaling path between User Agents A and B for the session established here. This Record-Route header field has also been passed to User Agent A by User Agent B within the response to the INVITE request. After receiving the Record-Route header field, it has created an internal route set for the further SIP signaling exchange with User Agent B. This route set contains the contact address of the SIP Proxy server according to the Record-Route header field. User Agent A must use the route set when sending the ACK message for User Agent B. It sends the message ACK in step (8) to the Proxy server. User Agent A transfers its route set entry (here: 89.89.89.89) in the SIP header field Route within the ACK message (8). The Proxy server receiving the ACK message deletes the route set entry concerning itself, adds a Via header field to the message, and forwards it to the temporary SIP URI of User Agent B given as Request URI in the Start Line (9).

The now existing session is terminated by subscriber B in step (10) by sending the SIP message BYE. Based on the Record-Route header field included in the INVITE message, User Agent B has also created a route set for further SIP signaling exchange with User Agent A. Due to the presence of this route set, User Agent B does not send the BYE message to User Agent A's temporary SIP URI given as Request URI in the start line but to the Proxy server. Within the BYE message, User Agent B transfers its route set entry (here: 89.89.89.89) in the Route header field. The Proxy server receiving the BYE message in step (10) deletes the route set entry concerning itself, adds a Via header field to the message, and forwards it to the temporary SIP URI of User Agent A given in the Start Line as Request URI (11). User Agent A confirms the BYE message in step (12) with the status information 200 OK, which it sends to the contact address of the Proxy server specified in the uppermost Via header field of the BYE message. This makes it clear that SIP responses are always sent back the same way that the corresponding SIP requests were sent initially. As a result, the Proxy server deletes the relevant Via header field from the status information and forwards it in step (13) to User Agent B, whose contact address it reads from the remaining Via header field.

In this context, it should also be mentioned that typically all SIP contact information includes not only an IP address but also a port number. The default port is 5060. In the SIP routing example shown in Figure 1.23 and explained above, we omitted the port numbers for simplicity [173].

The SDP (Session Description Protocol) is applied in the context of SIP for media description in multimedia communication. Just like SIP, it was standardized by the IETF in RFC 4566 [9], whereby the SDP messages are also purely text-based. SDP is used to exchange media types (audio, video, etc.) as well as contact parameters (IP address and port number) and an enumeration of the codecs available per medium on the particular end device (e.g., G.711, G.723, etc. for voice) between the SIP User Agents or gateways.

The SIP message shown in Figure 1.20 already contained SDP in the Message Body, described by SDP parameters. Since SDP is an independent protocol, it is also characterized by numerous parameters, only a few of which are essential in the context of SIP. These are listed in Table 1.4, including their functions.

Tab. 1.4: SDP parameters relevant for SIP, their functions, and examples

SDP parameter	Function	Example
c (Connection Data)	IP receiving address for user data	c=IN IP4 192.104.123.234
m (Media Descriptions)	Specification of a medium that is to be part of a media session: Media type (e.g., audio or video), receiving port, transport protocol for user data (e.g., RTP/AVP (Audio Video Profile)), supported codecs in the form of PTs (Payload Type number [6]) in the desired order	m=audio 2410 RTP/AVP 0 8 3 4 m=video 2412 RTP/AVP 34
a (Attributes)	With one or more attributes the m-parameter can be characterized in more detail	a=rtpmap:0 PCMU/8000 a=rtpmap:4 G723/8000 a=rtpmap:34 H263/90000 a=recvonly (receive only) a=sendrecv (send and receive)

The codec negotiation between the SIP User Agents takes place according to [4] following the offer/answer model. UA A offers, generally in the INVITE request, in the SDP A per m-parameter its codec request sequence as an enumeration (Offer), whereby the codec with the highest priority is on the far left: e.g., audio 34794 RTP/AVP 97 111 112 6 0 8 4 5 3 101.

UA B has three possibilities to respond with its SDP B, generally in the 200 OK response:

- Selection of a single codec from the offered list: audio 4474 RTP/AVP 0
- Repeat the received list omitting unsupported codecs (recommended): audio 4474 RTP/AVP 97 6 0 8
- Sending of an own independent codec list including at least one codec listed in the offer: audio 4474 RTP/AVP 9 7 0 10 (possible incompatibility).

In the recommended case (repetition) in the example, the audio codec with PT = 97 would be selected [173].

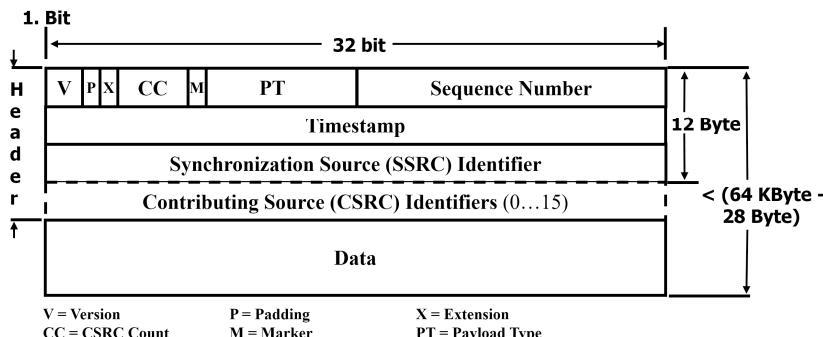
To conclude the brief explanation of SDP, Figure 1.24 shows the recording of an SDP message for audio and video with several codecs for selection.

```

Session Description Protocol
  Session Description Protocol Version (v): 0
  > Owner/Creator, Session Id (o): - 13100278460289054 2 IN IP4 10.10.10.101
    Session Name (s): X-Lite release 4.9.2 stamp 79048
    Connection Information (c): IN IP4 10.10.10.101
    Time Description, active time (t): 0 0
    Media Description, name and address (m): audio 53224 RTP/AVP 9 8 120 0 84 101
    Media Attribute (a): rtpmap:120 opus/48000/2
    Media Attribute (a): fmtp:120 useinbandfec=1; usedtx=1; maxaveragebitrate=64000
    Media Attribute (a): rtpmap:84 speex/16000
    Media Attribute (a): rtpmap:101 telephone-event/8000
    Media Attribute (a): fmtp:101 0-15
    Media Attribute (a): sendrecv
    Media Description, name and address (m): video 59046 RTP/AVP 34 115
    Media Attribute (a): rtpmap:34 H263/90000
    Media Attribute (a): fmtp:34 CIF=2;QCIF=2;VGA=2;CIF4=2
    Media Attribute (a): rtpmap:115 H263-1998/90000
    Media Attribute (a): fmtp:115 VGA=2;CIF=1;QCIF=1;CIF4=2;I=1;J=1;T=1
    Media Attribute (a): rtcp-fb:.* nack pli
    Media Attribute (a): sendrecv
  
```

Fig. 1.24: SDP message captured with protocol analysis software

After a successful SIP session setup with codec negotiated via SDP, the real-time user data transmission, e.g., for voice, takes place using RTP (Real-time Transport Protocol). It works connectionless end-to-end between User Agents or gateways and uses the connectionless layer 4 protocol UDP (User Datagram Protocol). RTP essentially provides as functionality the identification of the codec used, numbering of the transmitted RTP packets with increasing sequence numbers (+1 per successive RTP packet), and the transmission of a timestamp (+N per RTP packet, N = number of voice samples per RTP packet). There is a separate RTP session per medium and transmission direction. Figure 1.25 shows the structure of an RTP packet, and Figure 1.26 a corresponding protocol record from network practice [173; 5].

**Fig. 1.25:** Structure of an RTP packet

```

Real-Time Transport Protocol
> [Stream setup by SDP (frame 272)]
  10.. .... = Version: RFC 1889 Version (2)
  ..0. .... = Padding: False
  ...0 .... = Extension: False
  .... 0000 = Contributing source identifiers count: 0
  0... .... = Marker: False
  Payload type: ITU-T G.722 (9)
  Sequence number: 22993
  [Extended sequence number: 88529]
  Timestamp: 902683053
  Synchronization Source identifier: 0x66904236 (1720730166)
  Payload: dabb7fbed85cde7abaffbfff1ffdf5efad77edefeb5f4b2f6...

```

Fig. 1.26: RTP package captured with protocol analysis software

2 3G/4G Mobile Networks and NGN (Next Generation Networks)

2.1 3GPP Releases (3rd Generation Partnership Project)

Section 1.2 has already shown how the evolution of mobile networks from 2G to 3G (with the introduction of the NGN concept and Multimedia over IP with SIP) to 4G (with an All IP network) has taken place. We will now describe this development in more detail concerning the supported functionalities of the respective 3GPP release. Furthermore, the consistent development of 5G is shown. Table 2.1 illustrates this in an overview.

These development steps of the 3G (Release 99 to Release 7), 4G (Release 8 to Release 14) up to the 5G mobile networks (Release 15, 16, etc.) clearly show that there have always been revolutions besides evolution.

The first step, revolutionary from a network point of view, was Release 5. This was the first time that a network was standardized following the NGN concept (see Section 1.3). Also, a far-reaching decision was made in favor of SIP as the signaling protocol for Multimedia over IP (see Section 1.4), with the IMS (IP Multimedia Subsystem) as the SIP routing platform.

The second revolutionary step was the transition between 3G and 4G with Release 8. LTE (Long Term Evolution) is the first standardized radio access network technology that uses IP as the transport protocol for all services, including telephony. Together with the real-time capable EPC (Evolved Packet Core) and the IMS for SIP signaling, a completely standardized All IP mobile radio network was available for the first time, which also offered bit rates of up to 100 Mbit/s per radio cell.

The third and so far last revolutionary step in 3GPP mobile networks is Release 15 with the first phase of a 5G system [173; 57; 77; 63]. It should be noted that Table 2.1 lists only a few selected system features for 3GPP Releases 9 to 14 that are considered particularly important for evolution. Releases 15 to 18, we discuss in detail from chapter 4 onwards.

Tab. 2.1: 3GPP releases for 3rd, 4th and 5th generation mobile networks [57; 173; 77; 63]

Release 99	Release 4
<ul style="list-style-type: none">- 2000- Core network same as GSM + GPRS- Access network UTRAN + GERAN- Higher data rates, up to 2 Mbit/s- USIM (UMTS Subscriber Identity Module)- AMR Codec (Adaptive Multi-Rate), 3,4 kHz	<ul style="list-style-type: none">- 2001- Separation of signaling and user data in the core network- Instead of MSC MSC-Server + MGWs- CCS no.7 (Common Channel Signaling) over SIGTRAN- QoS architecture for PS Domain

Release 5	Release 6
<ul style="list-style-type: none"> – 2002 – NGN concept – Core network with IP Multimedia Subsystem (IMS) – Multimedia over IP with SIP – HSDPA (High Speed Downlink Packet Access), up to 14,4 Mbit/s downstream – Wideband AMR, 7 kHz 	<ul style="list-style-type: none"> – 2004 – MBMS (Multimedia Broadcast and Multicast Services) – WLAN/UMTS Interworking – IMS Phase 2 – Voice over IMS – HSUPA (High Speed Uplink Packet Access), up to 5,8 Mbit/s upstream
Release 7	Release 8
<ul style="list-style-type: none"> – 2007 – IMS enhancements for TISPAN NGN Release 1 and 2 as well as PacketCable – Emergency call via IMS – Voice Group Call Services (VGCS) for police, fire brigade, etc. – MIMO antenna technology (Multiple Input Multiple Output) – RAN enhancements: HSPA+ (High Speed Packet Access Plus), up to 42/22 Mbit/s down-/upstream 	<ul style="list-style-type: none"> – 2008 – SAE (System Architecture Evolution) for core network with EPC – eCall (vehicle emergency call) – Earthquake and tsunami warning – LTE for access network (E-UTRAN), up to 100/50 Mbit/s down-/upstream – Home NodeB/eNodeB – The basis for NGMN (Next Generation Mobile Networks)
Release 9	Release 10
<ul style="list-style-type: none"> – 2009 – Self-Organizing Networks (SON) 	<ul style="list-style-type: none"> – 2011 – Network optimization for M2M – Carrier Aggregation – LTE Advanced, up to 1000/500 Mbit/s down-/upstream
Release 11	Release 12
<ul style="list-style-type: none"> – 2012 – EVS Codec (Enhanced Voice Services) – WebRTC (Web Real-Time Communication between Browsers) with IMS – Proximity-based Services (ProSe) with Device-to-Device communication 	<ul style="list-style-type: none"> – 2015 – LTE-M (LTE for Machines) for M2M and IoT (Internet of Things)
Release 13	Release 14
<ul style="list-style-type: none"> – 2016 – NB-IoT (Narrowband-IoT) – Mission Critical Push To Talk (MCPTT) over LTE – LTE-Advanced Pro, up to 3/1,5 Gbit/s down-/upstream 	<ul style="list-style-type: none"> – 2017 – V2X (Vehicle to X) – Virtualization, Orchestration
Release 15	Release 16
<ul style="list-style-type: none"> – June 2019 – 5G Phase 1 	<ul style="list-style-type: none"> – December 2020 – 5G Phase 2
Release 17	Release 18
<ul style="list-style-type: none"> – December 2022 – 5G further development 	<ul style="list-style-type: none"> – Scheduled for mid 2024 – 5G-Advanced

2.2 IMS (IP Multimedia Subsystem) and NGN

As explained in Table 2.1 and already announced in Chapter 1, a significant development step in mobile networks is the transition to 3GPP Release 5 with the introduction of IMS, as shown in Figure 2.1, taking into account the NGN concept. A comparison with Release 4, as shown in Figure 1.15, illustrates the addition of IMS. Figure 2.1 shows the resulting network architecture with indicated IMS [29]. The core component of the IMS is the HSS (Home Subscriber Server). It is a database that, on the one hand, provides the HLR (Home Location Register) known from GSM/GPRS networks for mobility support, and on the other hand, contains the SIP user-profiles and offers the Location server functionality.

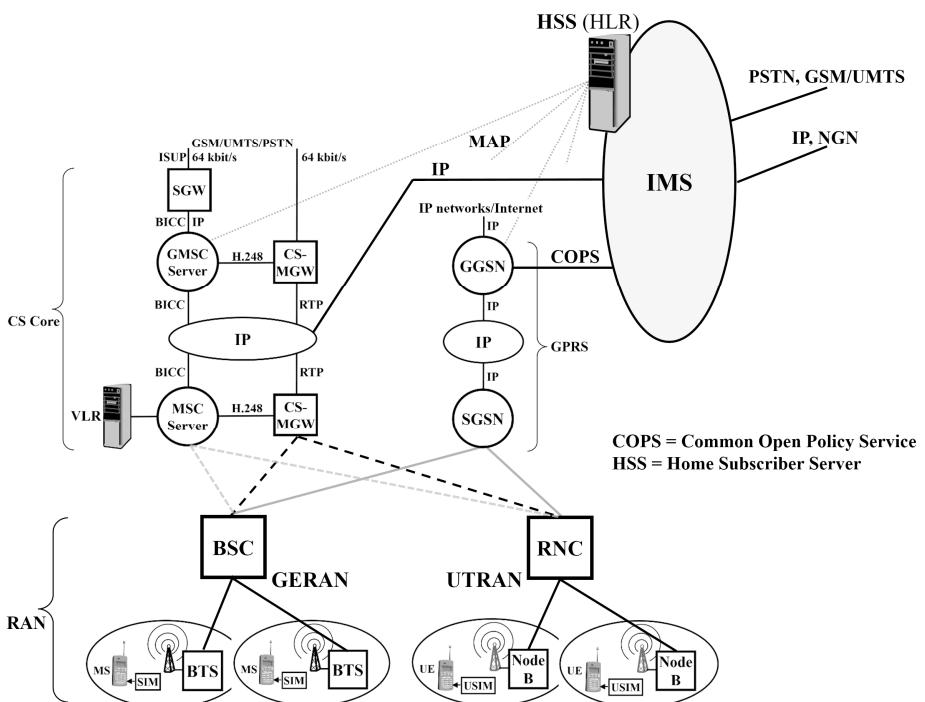


Fig. 2.1: 3GPP Release 5 mobile network

Figure 2.2 shows the internal structure of an IMS, and, in comparison with Figure 1.17, the reference to an NGN with SIP becomes obvious. The IMS is nothing else but the complete, comprehensive, and thoroughly standardized specification of the SIP routing platform for an NGN.

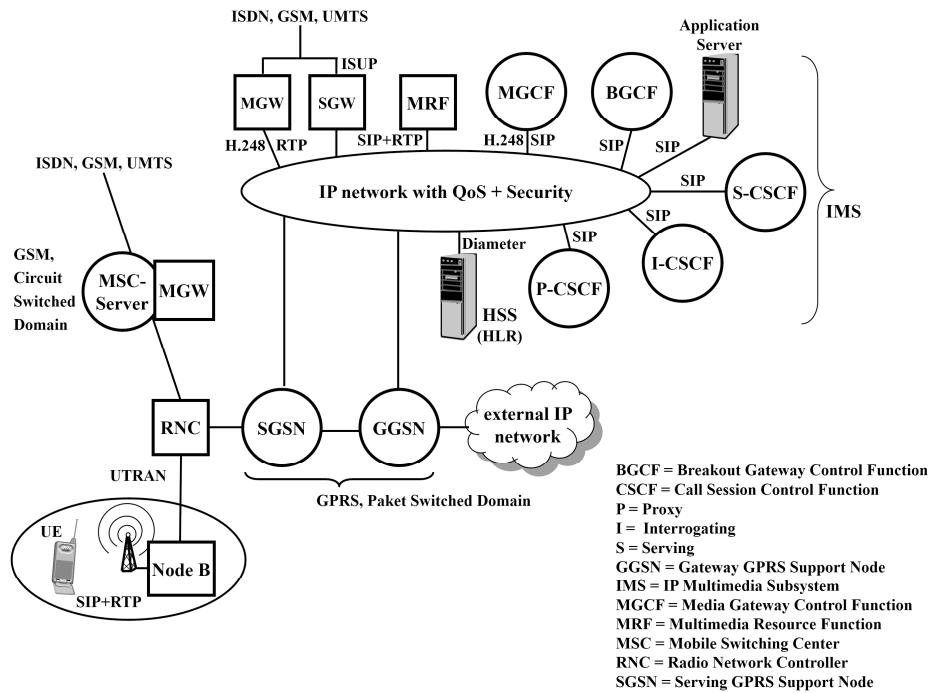


Fig. 2.2: IMS in a 3GPP Release 5 mobile network

The S-CSCF (Serving-Call Session Control Function) in Figure 2.2 mainly corresponds to the CS or a SIP Proxy/Registrar server in Figure 1.17. The S-CSCF, always located in the home network, registers the users and controls the SIP sessions, as well as the services and supplementary services. During registration, the HSS containing the user profiles, including the Location server, is queried. The S-CSCF communicates with the mobile devices, the User Equipment (UE), other CSCFs, and the application servers via SIP. The S-CSCFs are supported by optional I-CSCFs (Interrogating-CSCF). They serve as SIP contact points in the network, i.e., for all registration requests and all incoming connection requests from external sources. A corresponding I-CSCF determines which S-CSCF is responsible for querying the HSS. As a central contact, the I-CSCF ensures that the IMS network configuration is hidden from the outside. The border between GPRS or EPC and IMS is marked by a P-CSCF (Proxy-CSCF). Usually, the P-CSCF works exclusively as a proxy, i.e., SIP is not terminated; the messages are forwarded to an S-CSCF. The main reason for the three-part division of the CS into S-, I- and P-CSCF is mobility support (roaming). Every UE, but especially a UE in a visited network, needs a first contact point for SIP; this is the P-CSCF. However, the S-CSCF in the home network is always responsible for

SIP registration and SIP routing. Also, the I-CSCF implements the interface to external SIP/IP networks regarding signaling.

If a UE in Figure 2.2 requests a connection to a circuit-switched network, e.g., the ISDN or a GSM/UMTS network, the S-CSCF forwards this SIP request to the Breakout Gateway Control Function (BGCF) with SIP proxy server functionality. The BGCF routes the request to the BGCF of a neighboring network or selects the corresponding MGCF (Media Gateway Control Function, see MGC in Figure 1.17) in its network, which then controls the MGW (Media Gateway) accordingly. The Multimedia Resource Function (MRF) realizes, on the one hand, a conference server; on the other hand, multimedia data can be stored, evaluated, and generated, e.g., for speech recording, recognition, and synthesis [173].

Please note that the SIP network elements mentioned above are, first of all, only logical network elements. They can, therefore, be implemented independently or combined in one device. IMS implementations are available with separate servers or with a single server for P-, I-, and S-CSCF.

Due to the importance of IMS for connection-oriented communication, such as telephony in 5G networks, the network elements, protocols, and the functioning of IMS are discussed in more detail below. Concerning timeliness, the statements are based on 3GPP Release 8. Figure 2.3 gives a complete overview of the network elements, reference points, and protocols in IMS [30]. The IMS can be structured into four categories of logical network elements:

- Session Management and Routing: P-CSCF (Proxy-Call Session Control Function), I-CSCF (Interrogating-CSCF), S-CSCF (Serving-CSCF), E-CSCF (Emergency-CSCF), LRF (Location Retrieval Function)
- Databases: HSS (Home Subscriber Server), SLF (Subscription Locator Function)
- Interworking: IBCF (Interconnection Border Control Function), TrGW (Transition Gateway), BGCF (Breakout Gateway Control Function), MGCF (Media Gateway Control Function), IMS-MGW (IMS-Media Gateway)
- Services: AS (Application Server), MRFC (Multimedia Resource Function Controller), MRFP (Multimedia Resource Function Processor), MRB (Media Resource Broker).

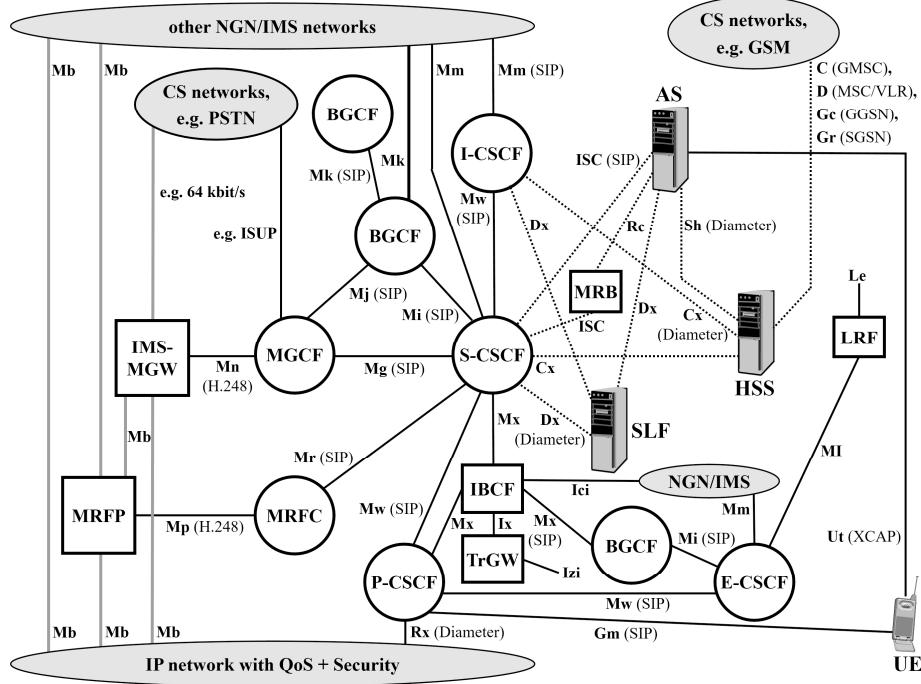


Fig. 2.3: Network elements, reference points, and protocols in IMS in 3GPP Release 8 [30]

The individual functions for all network elements mentioned and also shown in Figure 2.3 are briefly listed below [173; 30]:

- **S-CSCF:** represents SIP Registrar/Proxy server, registers users, stores registration information in the Location server, controls SIP connections, services, and supplementary services, communicates with UE, other CSCFs, and application servers, evaluates SDP. User profile data is loaded from HSS into the S-CSCF upon registration.
- **P-CSCF:** the first point of contact in the IMS for UE, usually works as a SIP proxy server, forms IPsec tunnels (IP security) to UEs, evaluates SDP for user access rights (media, codecs, etc.) and QoS (Quality of Service), cooperates with PCRF (Policy and Charging Rules Function) to provide the required QoS. Communicates with UE and S-CSCF or I-CSCF via SIP, with PCRF via Diameter protocol
- **I-CSCF:** at the interface to other IMSSs and IP multimedia networks, queries the HSS during registration for the responsible S-CSCF, usually works as a SIP proxy server. Communicates with S-CSCF or P-CSCF via SIP, with HSS and SLF via Diameter protocol

- E-CSCF (Emergency-CSCF) for the SIP routing of emergency calls, e.g., to the geographically nearest public safety answering point. If necessary, location information for a mobile device (UE) can be obtained via the LRF.
- LRF (Location Retrieval Function)
- HSS: central database with HLR and AuC (Authentication Center) functionality and the SIP user profiles with user identity, access rights, service trigger information for IMS. Access by MSC, GMSC of CS domain; SGSN, GGSN or EPC of PS domain; CSCF, ASs of IMS. Addressed via Diameter protocol
- SLF: database, offers I-CSCF, S-CSCF, and ASs the possibility to determine the address of the HSS responsible for a specific user. Access by I-CSCF, S-CSCF, and ASs. Addressed via Diameter protocol
- BGCF: decides where to exit the PSTN if necessary. Receives SIP request from S-CSCF when connecting to a circuit-switched network, selects MGCF in its own network, or routes request to BGCF in other networks. Communicates as SIP Proxy server with S-CSCF, MGCF, and other BGCFs using SIP
- MGCF: represents MGC. Protocol conversion ISUP/SIP or BICC/SIP. Controls IMS-MGW via H.248/Megaco protocol. Communicates with S-CSCF or BGCF via SIP, with CS network via ISUP or BICC, with IM-MGWs via H.248
- IMS-MGW: MGW for user data conversion, e.g., RTP/64 kbit/s. Controlled by MGCF via H.248 protocol. Generation of call progress tones and announcements, if necessary, provision of transcoding. Communicates over an IP transport network using RTP, with the PSTN based on 64 kbit/s channels and with the MGCF using H.248
- IBCF (Interconnection Border Control Function) as Session Border Controller for signaling (SBC-S) at the transition to other NGNs
- TrGW (Transition Gateway): gateway for NAPT (Network Address and Port Translation) and IPv4/IPv6 protocol conversion in the media path. Controlled by an IBCF
- MRB: supports the use of a common pool of different MRF resources (Media Resource Function, MRFC + MRFP) in interaction with S-CSCF and application servers. The MRB allocates MRF resources to sessions for a specific application, e.g., due to available capacities or required QoS.
- AS: for the provision of services, especially value added services. S-CSCF routes SIP requests/responses to a specific AS based on internal filter criteria or filter criteria queried by HSS. Communicates with S-CSCF via the so-called ISC interface (IMS Service Control) using SIP, with data servers, e.g., using HTTP (Hyper-text Transfer Protocol), with HSS or SLF via Diameter protocol, with a UE, e.g., using XCAP (XML Configuration Access Protocol)
- MRFC: for the control of user data processing in the IMS. Controlled via S-CSCF using SIP, i.e., MRFC represents a SIP UA function. Communicates with S-CSCF using SIP, controls MRFP using H.248 protocol

- MRFP: for user data handling in the IMS, such as voice recording and playback, video recording and playback, speech recognition, conversion of text to speech, multimedia conferences, and transcoding of multimedia data.

Compared to IETF standard SIP (see Section 1.4), there are extensions for IMS and mobility-specific support in the 3GPP releases that include SIP. The most relevant ones are summarized below [173]:

- An additional Private User ID was introduced for SIP IDs. It is stored on the SIM card and identifies the user's service subscription or user profile in the HSS. This Private User ID is only used for authentication during the registration process, not for SIP routing. Also, there are the permanent SIP URIs common in SIP, which are called Public User IDs here. Each user is assigned one Private and N Public User ID.
- The authentication is carried out in extension to IETF-SIP via AKA (Authentication and Key Agreement) within the otherwise usual SIP digest procedure.
- To identify the currently visited network, the SIP header field P-Visited-Network-ID is used.
- The P-Charging Vector provides charging information of the different SIP network elements.
- Path: With this header field, a P-CSCF informs the S-CSCF during the registration of a UE (via REGISTER request) that the UE uses this special P-CSCF for SIP signaling. The S-CSCF stores the path, the used P-CSCF. This information is then entered into the Route header field of the INVITE request in case of an incoming call to ensure that the P-CSCF, initially selected by the UE, is passed through. This is mandatory because communication between P-CSCF and UE is performed via an IPsec tunnel for security reasons.
- Service-Route: With this header field, an S-CSCF informs a UE about the used S-CSCF within its response 200 OK to the registration request. The UE stores the Service route. Based on this information, only P-CSCF and the now directly addressable S-CSCF are passed through when a session is set up, no longer the I-CSCF that was also included in the registration.
- Support of the SIP request PRACK as a temporary ACK to “stop” a session until a precondition is met, e.g., that the requested QoS can be provided end-to-end by the network.

The above comments on IMS are supplemented by considerations regarding a registration process and session setup using IMS in a 4G/LTE mobile network [41].

Figure 2.4 shows the network architecture focusing on the IMS registration process. Numbers indicate the sequence in which the REGISTER message passes through the various SIP network elements. It is also specified when which database – DNS to determine the responsible I-CSCF, HSS to identify the relevant S-CSCF – is queried.

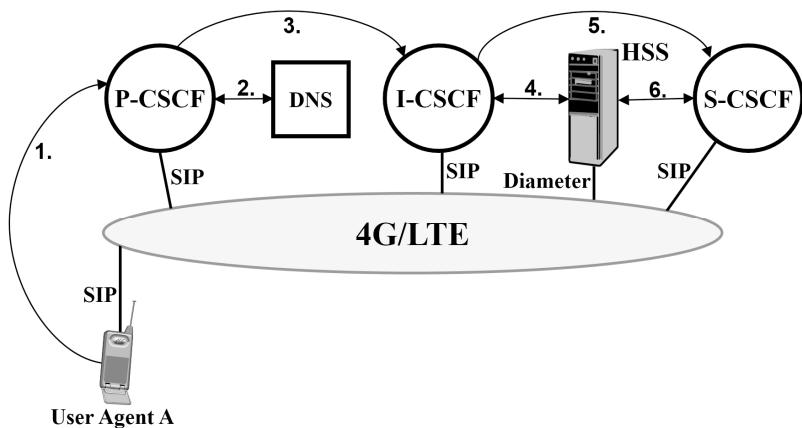


Fig. 2.4: Network architecture for SIP registration in IMS

Figure 2.5 describes the procedure for SIP registration in IMS. The same numbers represent the sequence shown in Figure 2.4.

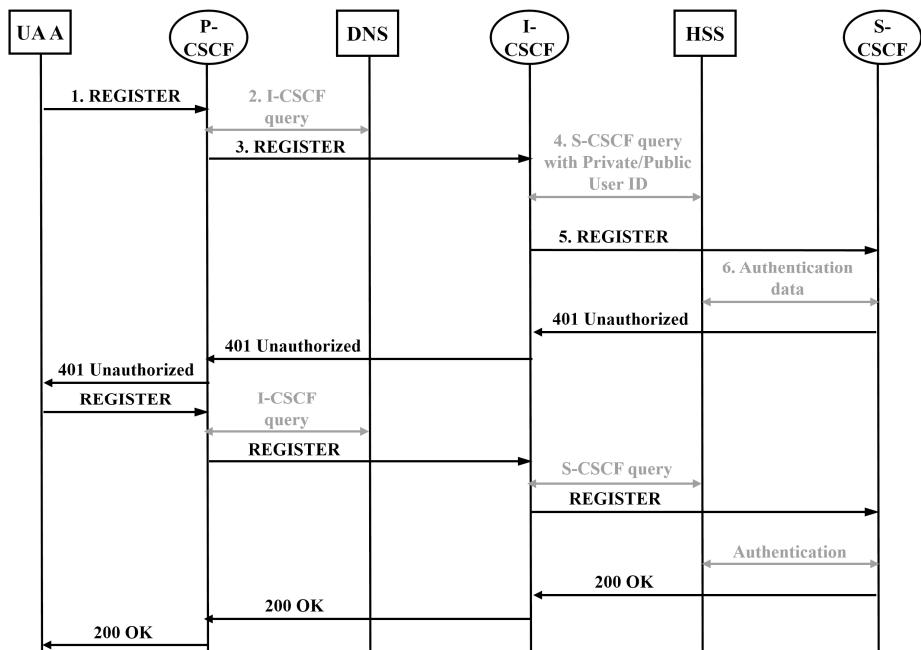


Fig. 2.5: Procedure of a SIP registration in the IMS

Figure 2.6 then shows the network architecture focusing on a session setup in the IMS. The sequence in which the INVITE message passes through the various SIP network elements or when the I-CSCF queries the HSS to determine the responsible S-CSCF is indicated with numbers. Figure 2.7 also shows the complete procedure for establishing a SIP session in the IMS, using the numbers from Figure 2.6 to identify the sequence.

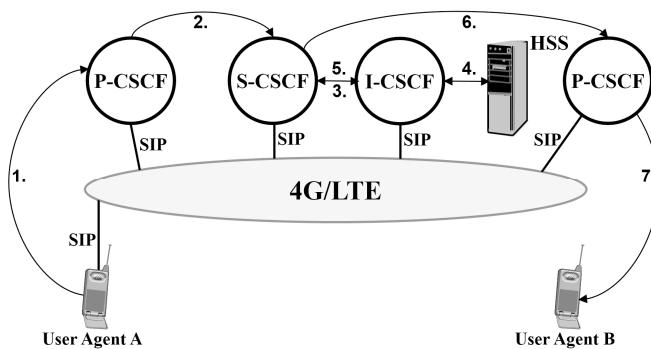


Fig. 2.6: Network architecture for a SIP session setup in the IMS

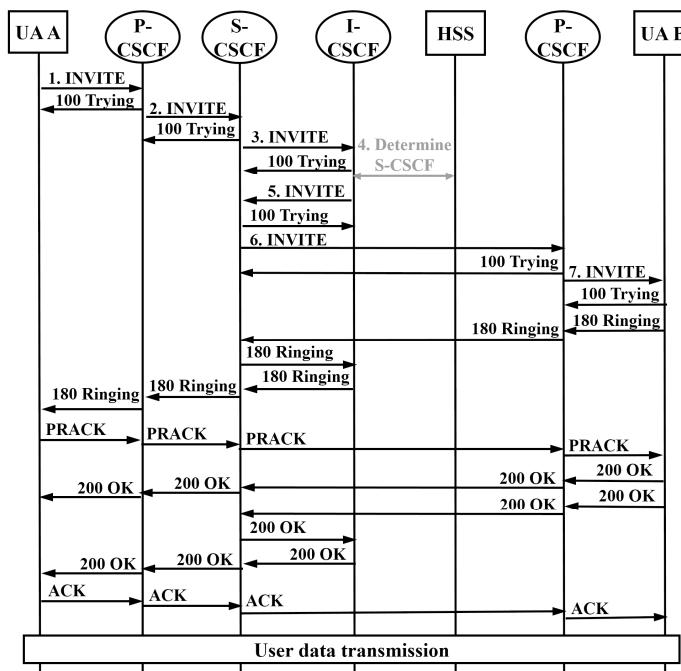


Fig. 2.7: Procedure for establishing a SIP session in the IMS

Finally, Figure 2.8 shows for the IMS which SIP network elements are involved in a SIP session setup in the case of roaming. It is noticeable that in a visited network, only the P-CSCF is part of the SIP path [173; 154].

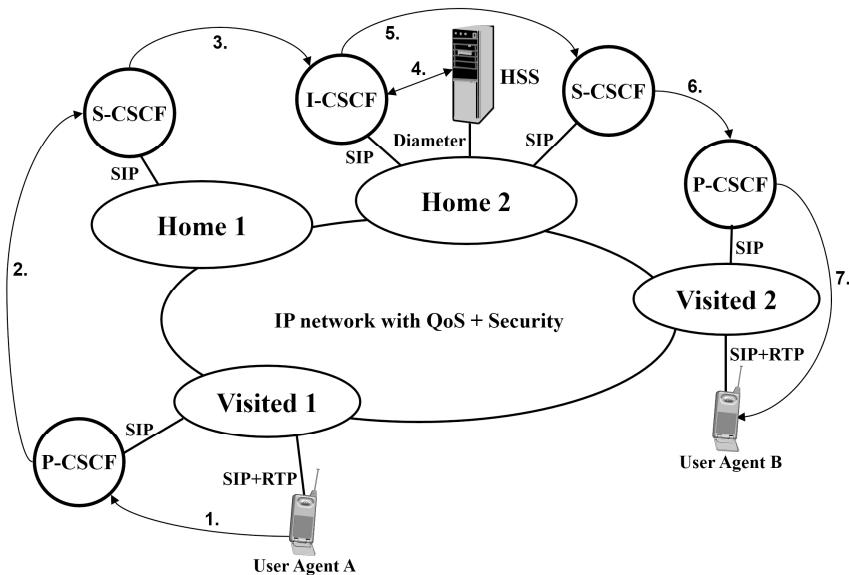


Fig. 2.8: Network architecture for roaming in IMS-based networks

2.3 H.248/Megaco Protocol

Figure 2.3 shows that three protocols mainly dominate an IMS: first of all, of course, SIP, including SDP for signaling, but then also Diameter for database access and the H.248 protocol for controlling media gateways (IMS-MGW) or general network elements processing user data (MRFP). We discuss H.248 in more detail in this section.

The H.248 protocol, also called Megaco, was initially specified jointly by the ITU-T and the IETF. Meanwhile, the responsibility for the standardization of H.248 is entirely with ITU-T; the latest H.248 standard is version 3 [104].

In IMS, as shown in Figure 2.3, the H.248 protocol is used between MGCF and IMS-MGW to implement a decomposed gateway and between MRFC and MRFP to implement a media resource function (MRF). The H.248 protocol operates in master (MGCF, MRFC) slave (IMS-MGW, MRFP) mode. UDP, TCP, or also SCTP (Stream Control Transmission Protocol) can be used as a transport protocol. The H.248 messages can be formatted text-based or binary ASN.1-coded (Abstract Syntax Notation) [104].

The H.248 standard is based on a connection model. This describes objects within an MGW that can be controlled by an MGC. A connection within an MGW consists

of endpoints (terminations) – sources and sinks of a media transmission – and an associated context that describes the relationships (associations) between the terminations. A difference is made between temporary (ephemeral), e.g., for RTP/IP stream, and physical, e.g., for 64 kbps channel, terminations. Figure 2.9 shows some examples of H.248 contexts.

Context 1 describes a user data connection of a conference between a VoIP user (RTP stream) and two subscribers in the SCN bearer channel, switched circuit network). The X represents the communication relationship between the terminations. Context 2 shows a simple NGN-SCN user data connection for a similar application. In context 2, a VoIP connection is maintained (parked) for assignment to another termination. Finally, the Null Context shown in Figure 2.9 indicates a termination that is currently not included in a context, i.e., is in the idle state [104; 162].

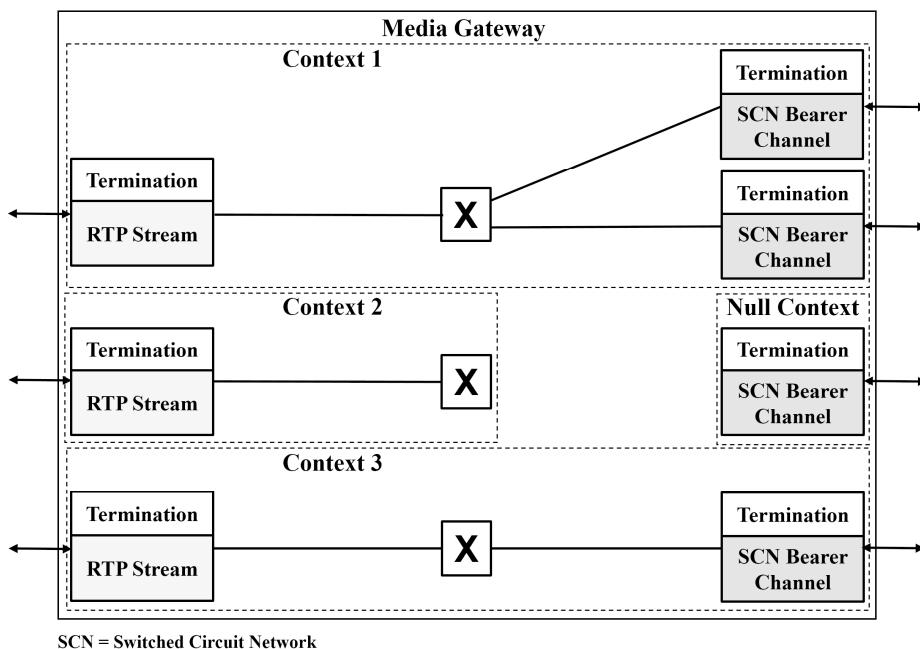


Fig. 2.9: Example of an H.248 connection model with contexts [104; 162]

The objects (terminations, contexts) of an MGW are controlled by an MGC using H.248 request and reply messages within transactions. Figure 2.10 shows the basic structure of such an H.248 message. It consists of a header and one or more transactions. A distinction is made between a request and a resulting reply. If the transaction cannot yet be completed, the request sender is informed of this with a pending

transaction. Each transaction consists of one or more actions, whereby an action does not have its own ID but acts as a placeholder for one or more commands. The Commands are used to create, change, query, or delete the contexts and terminations. Table 2.2 provides an overview [104; 170].

H.248 Message

Header: Version, MediagatewayID (IP-Address:Port-Number)

Transaction 1: Transaction (Request or Reply or Pending, ID, Context)

Action 1

Command 1

Termination(s)

Descriptors

:

Command x

Termination(s)

Descriptors

:

Action y

Command 1

Termination(s)

Descriptors

:

:

Transaction z

Fig. 2.10: Basic structure of an H.248 message [104; 170]

Tab. 2.2: H.248 commands [104]

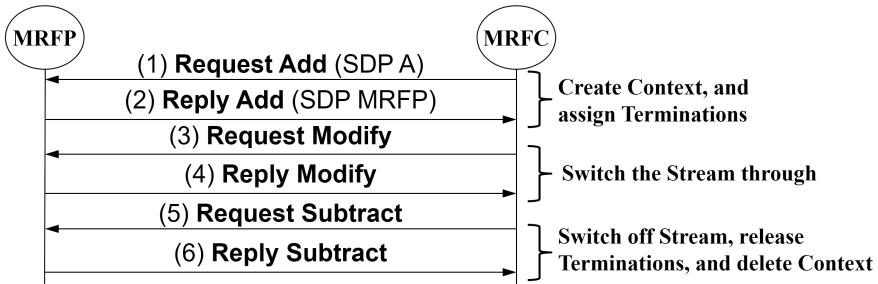
Command	Function
Add	Adds a termination to a context. The first Add command creates a context.
Modify	Modifies the properties of a termination
Subtract	Disconnects a termination from a context. The last subtract command for a context deletes it. The statistical information collected so far is sent to the MGC.
Move	Moves a termination to another context
AuditValue	Returns the current state of properties of a termination
AuditCapability	Returns all possible parameter values for a termination
Notify	An MGW informs an MGC about the occurrence of events, e.g., DTMF tones (Dual Tone Multi-Frequency).
ServiceChange	An MGW informs an MGC that an MGW is available (registration) or that termination or group of terminations is out-of-service or in-service. Accordingly, an MGC can modify terminations in an MGW.

Each command in Figure 2.10 refers to one or more terminations (media source and/or sink) and is parameterized by so-called descriptors. Table 2.3 gives an overview of selected descriptors [104].

Tab.2.3: Selected descriptors to describe terminations [104; 162]

Descriptor	Function
Media	Defines stream ID, specifies properties of a single media stream. The list of Local Control, Local and/or Remote Descriptors for a single stream can be inserted here. Several media descriptors together can describe a multimedia stream.
Stream	Describes properties for individual bidirectional streams, including transmit and/or receive properties such as send only, receive only, send and receive. Contains this information in a list of Local Control, Local and/or Remote descriptors
Local	Describes properties of the media flow received by the MGW. With the text-based H.248 protocol, SDP (see Section 1.4) is used here for the media description.
Remote	Describes properties of the media flow that the MGW sends to the remote communication partner. With the text-based H.248 protocol, SDP (see Section 1.4) is used for the media description.
LocalControl	Specifies properties describable in packages that are of interest to both MGW and MGC
Events	Describes possible events that can occur at the MGW and corresponding subsequent actions. For example, DTMF tones can be detected.
Signals	Describes signals that can be assigned to terminations. E.g., a dial tone, free tone, or busy tone, can be applied to an interface.
Audit	Specifies requested audit information
Statistics	Statistical information on terminations or streams in Subtract or Audit commands

Based on the above explanations of the H.248 protocol, the exemplary message flow shown in Figure 2.11 can be understood. It describes the situation in an IMS where an MRFC requests an MRFP to switch through an audio stream for a user A. This is done with three contexts in three transactions, each with request and reply, with the commands Add, Modify, and Subtract. For a better understanding and the necessary practical orientation, Figures 2.12 to 2.15 show the concrete H.248 messages (1), (2), (3), and (5) from Figure 2.11.

**Fig. 2.11:** Exemplary H.248 message flow for providing a media stream for user A

```

Internet Protocol Version 4, Src: 10.94.8.51, Dst: 10.94.9.115
Stream Control Transmission Protocol, Src Port: 2944 (2944), Dst Port: 2944 (2944)
MEGACO
Start token: !
Version: 2
MediagatewayID: [10.94.8.51]:2944
Transaction: Request
Transaction ID: 7566
Context: Choose one
▼ Command: Priority
  Command: Priority
  Priority: 6
▼ Command: Add
  Command: Add
  Termination ID: rtp/1/$
▼ Descriptors
  ▼ Media Descriptor
    StreamID: 1
    ▶ Local Control Descriptor
    ▶ Local Descriptor
  ▼ Local Descriptor
    ▼ Session Description Protocol
      Session Description Protocol Version (v): 0
      ▶ Connection Information (c): IN IP4 10.94.9.10
      ▶ Media Description, name and address (m): audio 45188 RTP/AVP 116 118 111 110
      ▶ Bandwidth Information (b): AS:49
      ▶ Bandwidth Information (b): RS:612
      ▶ Bandwidth Information (b): RR:1837
      ▶ Media Attribute (a): rtpmap:116 AMR-WB/16000/1
      ▶ Media Attribute (a): rtpmap:118 AMR/8000/1
      ▶ Media Attribute (a): rtpmap:111 telephone-event/16000
      ▶ Media Attribute (a): rtpmap:110 telephone-event/8000
      ▶ Media Attribute (a): fmp:116 mode-change-capability=2; max-red=0
      ▶ Media Attribute (a): fmp:118 mode-change-capability=2; max-red=0
      ▶ Media Attribute (a): fmp:111 0-15
      ▶ Media Attribute (a): fmp:110 0-15
      ▶ Media Attribute (a): pttime:20
  ▼ Events Descriptor
    RequestID: 1
    pkgdName: G/CAUSE

```

Fig. 2.12: H.248 text-based request message (1) with command Add captured by protocol analysis software

```

Internet Protocol Version 4, Src: 10.94.9.115, Dst: 10.94.8.51
Stream Control Transmission Protocol, Src Port: 2944 (2944), Dst Port: 2944 (2944)
MEGACO
Start token: !
Version: 2
MediagatewayID: [10.94.9.115]:2944
Transaction: Reply
Transaction ID: 7566
Context: 1004929
	Command: Add
		Command: Add
		Termination ID: rtp/1/1016919
	Descriptors
		Media Descriptor
			StreamID: 1
		Local Descriptor
			Session Description Protocol
				Session Description Protocol Version (v): 0
				> Connection Information (c): IN IP4 10.94.9.14
				> Media Description, name and address (m): audio 20086 RTP/AVP 116 111
				> Bandwidth Information (b): AS:41
				> Media Attribute (a): rtpmap:116 AMR-WB/16000
				> Media Attribute (a): rtpmap:111 telephone-event/16000
				> Media Attribute (a): fmtcp:116 max-red=0; mode-change-capability=2
				> Media Attribute (a): fmtcp:111 0-15
				> Media Attribute (a): ptimc:20
				> Media Attribute (a): maxptime:40

```

Fig. 2.13: H.248 text-based reply message (2) with command Add captured by protocol analysis software

```

Internet Protocol Version 4, Src: 10.94.8.51, Dst: 10.94.9.115
Stream Control Transmission Protocol, Src Port: 2944 (2944), Dst Port: 2944 (2944)
MEGACO
Start token: !
Version: 2
MediagatewayID: [10.94.8.51]:2944
Transaction: Request
Transaction ID: 7568
Context: 1004929
	Command: Priority
		Command: Priority
		Priority: 9
	Command: Modify
		Command: Modify
		Termination ID: rtp/1/1016919
	Descriptors
		Signal Descriptor
			pkgdName: AN/APF
				st=1,nc={to,or},AN=de_990,NOC=0

```

Fig. 2.14: H.248 text-based request message (3) with command Modify captured by protocol analysis software

```

Internet Protocol Version 4, Src: 10.94.8.51, Dst: 10.94.9.115
Stream Control Transmission Protocol, Src Port: 2944 (2944), Dst Port: 2944 (2944)
MEGACO
Start token: !
Version: 2
MediagatewayID: [10.94.8.51]:2944
Transaction: Request
Transaction ID: 7592
Context: 1004929
	Command: Priority
		Command: Priority
		Priority: 9
	Command: Subtract
		Wildcarded response to a command
		Command: Subtract
		Termination ID: WildCard all
	Descriptors
		Audit Descriptor

```

Fig. 2.15: H.248 text-based request message (5) with command Subtract captured by protocol analysis software

2.4 Diameter Protocol

As already mentioned at the beginning of Section 2.3, regarding Figure 2.3, the Diameter protocol plays a significant role in IMS-based networks besides SIP and H.248. It is used to exchange AAA information (Authentication, Authorization and Accounting). It is a further development of the RADIUS protocol (Remote Authentication Dial In User Service), especially concerning expandability and flexibility. In the IMS, or more generally in an NGN, Diameter is applied for AAA-related communication between servers and databases (see Figure 2.3):

- S-CSCF – HSS
- S-CSCF – SLF
- I-CSCF – HSS
- I-CSCF – SLF
- AS – HSS.

Essential for the flexibility of the application and the expandability is the split of the protocol specification into a Diameter Base Protocol according to RFC 6733 [12] for the elementary functions and various protocol extensions adapted to the supported applications, so-called Diameter Applications. In the IMS context, the Diameter SIP Application in RFC 4740 [11] should be mentioned, which will be referred to later.

The Diameter Base protocol provides the following functionalities:

- Diameter message exchange with the transmission of AVPs (Attribute-Value Pair)

- All Diameter data are described in the form of AVPs. An AVP consists of a header and a data field and encapsulates the AAA and/or Diameter routing information.
- Negotiation of the supported properties
- Error notifications
- Expandability by adding new Diameter Applications, protocol commands, and/or AVPs
- Transport of user authentication data
- Transport of service-specific authorization data
- Exchange of data on resource use for accounting or capacity planning purposes
- Forwarding and routing of Diameter messages in a server hierarchy.

Diameter is an extended client-server protocol. The client, for example, an S-CSCF, requests a server, for example, the HSS, to provide an AAA function for a user or service. Extended here means that peer-to-peer communication is also supported, i.e., a Diameter server can also initiate communication with a Diameter client. According to [12], the transport protocols for Diameter can be TCP or SCTP (Stream Control Transmission Protocol). In both cases, the default port number is 3868.

As already noted, a Diameter message consists of a header followed by data encapsulated in AVPs. Figure 2.16 shows the header structure of a Diameter message. As is often the case with Internet protocols, the header fields are arranged in a 32-bit row structure. After the version, 1, follows a length declaration for the entire message, including the attached AVPs in a multiple of 4 Byte (32 bit). This is followed by several flags, including R (1 = Request, 0 = Answer), P (1 = Proxiable: message may be forwarded), E (1 = Error: message contains protocol errors), and a 3-Byte command code. The latter specifies the actual function of the Diameter message. The header is supplemented by an Application-ID, which marks the affiliation of the message to a certain Diameter application, a Hop-by-Hop Identifier, which is identical for related requests and answers, as well as an End-to-End Identifier, which can be used to detect unrequested message duplicates [12].

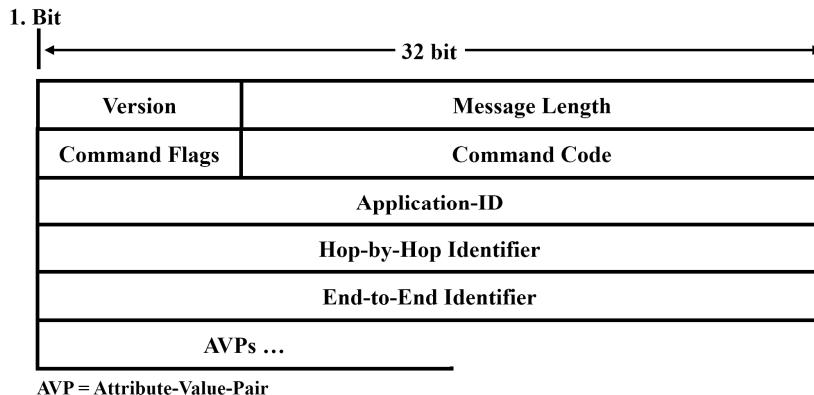


Fig. 2.16: Header of a Diameter message [12]

A Diameter command with the subtype Request or Answer (indicated by the R flag), represented by a 3-digit decimal code, is described by an identifier with a 3-letter abbreviation: e.g., 257, Capabilities Exchange Request (CER) [12].

As already mentioned, the actual AAA data, which are part of each Diameter message, are described in so-called AVPs. Figure 2.17 shows the structure of such an AVP. The three-digit decimal codes defining the diameter-specific AVPs are assigned by the IANA (Internet Assigned Numbers Authority), starting from the value 257. 1 to 256 are reserved for RADIUS to ensure backward compatibility. The AVP code, together with the Vendor ID field, uniquely identifies a Diameter attribute, i.e., the requested or delivered Diameter data. Several flags follow the AVP code, including V (1 = vendor-specific: optional vendor ID field available) and M (1 = mandatory: AVP must be supported). VP Length specifies the length of an AVP, including the data in Byte. The data field finally contains the actual AAA data of this AVP [12].

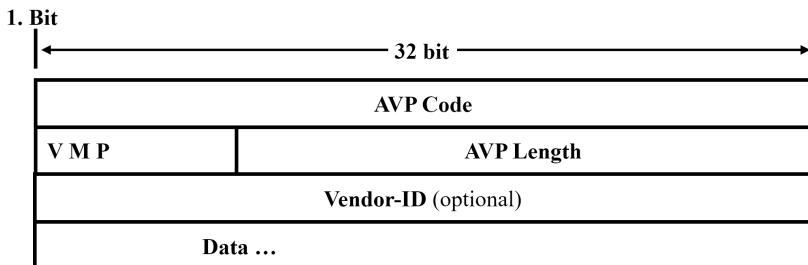


Fig. 2.17: AVP header [12]

Table 2.4 gives an overview of selected Diameter Base Protocol AVPs and their functions. This is relevant for the examples from IMS practice considered later.

Tab. 2.4: Selected Diameter Base Protocol AVPs and their functions [12]

Attribute name	AVP Code	Function
Auth-Application-Id	258	Indicates the support of authentication and authorization of an application
Auth-Session-State	277	Specifies whether the state of a Diameter session is maintained
Destination-Realm	283	Identifies area/domain of the destination of a Diameter message
Experimental-Result	297	Indicates whether a vendor-specific query was completed successfully or whether an error occurred
Origin-Host	264	Identifies the author of the Diameter message
Origin-Realm	296	Identifies area/domain of the originator of a Diameter message
Result-Code	268	Returns the result of a request with a result code in the form 1xxx (notifying), 2xxx (successful), or 3xxx to 5xxx (error)
Session-Id	263	Identifies a Diameter session. All Diameter messages of the same session contain the same session ID.
User-Name	1	Contains the name of the considered user
Vendor-Id	266	ID assigned by IANA to a corresponding software vendor
Vendor-Specific-Application-Id	260	Indicates the support of a vendor-specific application

Particularly important for the IMS and Diameter is the Diameter Session Initiation Protocol (SIP) Application, according to RFC 4740 [11], which extends the Diameter Base Protocol described above. It is used to authenticate SIP users and to authorize the use of resources within SIP sessions for multimedia services. The Diameter server can also send updated user profiles to a SIP server. Further, information can be provided to a SIP server to locate other SIP servers. In consideration of these functionalities provided by the Diameter protocol for SIP/IP-based networks, the SIP servers I-CSCF, S-CSCF, and an AS each contain a Diameter client. HSS and SLF represent associated Diameter servers.

Figure 2.18 shows a Diameter-using SIP network architecture, according to [11]. In comparison with Figure 2.3, SIP server 1 is an I-CSCF whose main task is to find the SIP server 2 responsible for a specific SIP UA by Diameter. Server 2 represents the S-CSCF in Figure 2.3. It provides the authentication and authorization of a user or a UA incl. SIP registration and routing by access to a Diameter server. This is, according to Figure 2.3, the HSS. The Diameter SL (Subscriber Locator), an SLF from

Figure 2.3, acts as a Diameter Redirect server and, on request, provides the Diameter server responsible for a particular user, i.e., here, the HSS.

Figure 2.18 shows not only the SIP and Diameter network elements but also the essential Diameter messages exchanged between them, represented by the commands listed in Table 2.5.

Tab. 2.5: Selected Diameter commands for SIP-based communication [11]

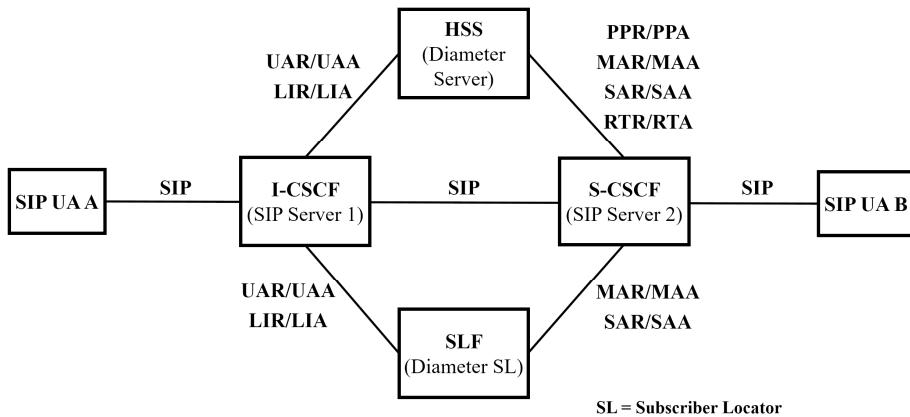
Diameter command	Command code	Function
User-Authorization-Request (UAR)	283 (300)	Request from SIP server 1 (I-CSCF) to Diameter server (HSS), which SIP server 2 (S-CSCF) is responsible for the user to be registered
User-Authorization-Answer (UAA)	283	Provides responsible SIP server 2 (S-CSCF) and thus authorizes SIP registration
Multimedia-Auth-Request (MAR)	286 (303)	SIP server 2 (S-CSCF) requests authentication and authorization for a user's SIP service from Diameter server (HSS)
Multimedia-Auth-Answer (MAA)	286	Result of the authentication and authorization process
Server-Assignment-Request (SAR)	284 (301)	SIP server 2 (S-CSCF) informs Diameter server (HSS) that SIP server 2 (S-CSCF) has completed the authentication process
Server-Assignment-Answer (SAA)	284	Diameter server (HSS) delivers user profile to SIP server 2 (S-CSCF)
Location-Info-Request (LIR)	285 (302)	Request from SIP server 1 (I-CSCF) to Diameter server (HSS), via which SIP URI the SIP server 2 (S-CSCF) responsible for the user can be reached (SIP routing information)
Location-Info-Answer (LIA)	285	Diameter server (HSS) returns SIP URI of the SIP server 2 (S-CSCF) responsible for the user

The additional command codes in brackets in Table 2.5 are applied according to RFC 3589 [7], especially in 3GPP Release 5, and are used in the following practical examples.

Table 2.6 shows a selection of AVPs used in addition to those in Table 2.4. Again, special AVP codes used in the corresponding Release 5 system are shown in brackets.

Tab. 2.6: Selected Diameter SIP application AVPs and their functions [11; 10]

Attribute name	AVP code	Function
SIP-Accounting-Information	368 (618, Charging-Information)	Contains parameter addresses of network elements that can collect accounting information
SIP-AOR	122 (601, Public-Identity)	Contains permanent SIP URI of the user
SIP-Server-URI	371 (602, Server-Name)	SIP URI to identify a SIP server
SIP-Server-Capabilities	372 (603)	Requirements for selecting a suitable SIP server
SIP-Server-Assignment-Type	375 (614)	Specifies the type of SIP server access required, for example, registration
SIP-Auth-Data-Item	376 (612)	Contains SIP authentication and/or authorization information
SIP-Number-Auth-Items	382 (607)	Number of SIP authentication and/or authorization credentials
SIP-Visited-Network-Id	386 (600)	Identifies a visited network
SIP-User-Data	389 (606)	User profile
SIP-User-Data-Already-Available	392 (624)	Indicates to the Diameter server whether the SIP server has received the required user profile

**Fig. 2.18:** SIP network architecture and Diameter protocol [11]

The relationships between SIP and the Diameter Protocol are worked out in the following based on the registration process for a UE or a SIP UA in the IMS, whereby the focus is on the network elements I-CSCF, S-CSCF, and HSS from Figures 2.3 and

2.18. Figure 2.19 shows the protocol procedures. The concrete Diameter protocol messages (see Table 2.5) from a practical example are shown in Figures 2.20 to 2.27. In this context, the most essential Diameter AVPs (see Tables 2.4 and 2.6) are presented in more detail.

As shown in Figure 2.19, the SIP proxy server I-CSCF (in the practical example with the IP address 10.0.2.110) receives a REGISTER request (1). The Diameter client of the I-CSCF then sends a User-Authorization-Request ((2) UAR, see Figure 2.20) to the Diameter server HSS (with the IP address 10.0.2.150) to request the responsible S-CSCF for the user to be now registered (AVP SIP-AOR or Public-Identity: `sip:bob@mnc001.mcc001.3gppnetwork.org`). The answer is provided by the User-Authorization-Answer ((3) UAA, see Figure 2.21), specifically with the AVP SIP-Server-Capabilities (server name: `sip:sccscf.mnc001.mcc001.3gppnetwork.org:5060`). Based on this response and a DNS query, the I-CSCF routes the SIP REGISTER request (4) to the S-CSCF (IP address 10.0.2.120). Subsequently, the Diameter client of the S-CSCF sends a Multimedia-Auth-Request ((5) MAR, see Figure 2.22) to the HSS for authentication and authorization of the registration (AVP SIP-Auth-Data-Item) for the user or their User Agent (AVP SIP-AOR or Public-Identity: `sip:bob@mnc001.mcc001.3gppnetwork.org`). The positive response here is indicated with Multimedia-Auth-Answer ((6) MAA, see Figure 2.23), especially in the AVP result code. The S-CSCF then sends the SIP Response 401 Unauthorized (7) to the I-CSCF, which forwards it to the P-CSCF or UA (8). After calculating the required authentication data in the UA, these are sent in a second SIP REGISTER message (9) and (10)) to the S-CSCF, which in turn verifies the authentication data and informs the HSS with Server-Assignment-Request ((11) SAR, see Figure 2.24) about the successful completion of the registration process with AVP SIP Server-Assignment-Type and requests the user profile with the AVP User-Data-Already-Available (here `USER_DATA_NOT_AVAILABLE`). The concrete user profile in XML format is then made available to S-CSCF in the response Server-Assignment-Answer ((12) SAA, see Figure 2.25) with the AVP SIP-User-Data or Cx-User-Data. The SIP Responses 200 OK ((13) and (14)) confirm the registration. Figure 2.19 also shows the case where the I-CSCF requests the responsible S-CSCF with Location-Info-Request ((15) LIR, see Figure 2.26) and receives the response (`sip:sccscf.mnc001.mcc001.3gppnetwork.org:5060`) with Location-Info-Answer ((16) LIA, see Figure 2.27) in the AVP server URI or server name.

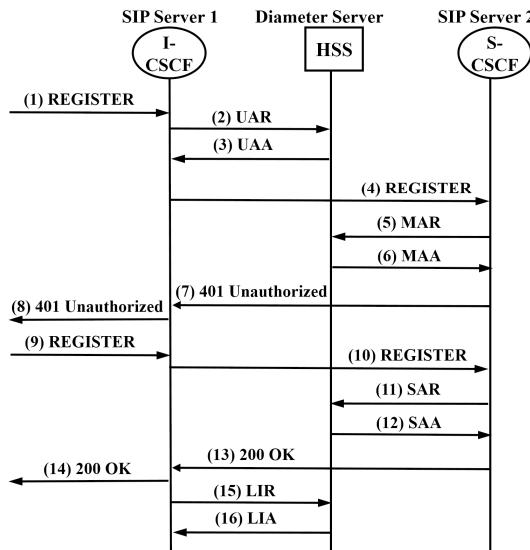


Fig. 2.19: SIP authentication by Diameter at registration

```

Internet Protocol Version 4, Src: 10.0.2.110, Dst: 10.0.2.150
Transmission Control Protocol, Src Port: 3868, Dst Port: 49556, Seq: 1, Ack: 1, Len: 388
Diameter Protocol
  Version: 0x01
  Length: 388
  Flags: 0xc0, Request, Proxyable
    1... .... = Request: Set
    .1... .... = Proxyable: Set
    ..0.... .... = Error: Not set
    ...0.... .... = T(Potentially re-transmitted message): Not set
    ....0... .... = Reserved: Not set
    ....0.. .... = Reserved: Not set
    ....0..0.... = Reserved: Not set
    ....0...0.... = Reserved: Not set
  Command Code: 300 User-Authorization
  ApplicationId: 3GPP Cx (16777216)
  Hop-by-Hop Identifier: 0x3d493cba
  End-to-End Identifier: 0x4ccdf17e
  [Answer_In: 21]
  > AVP: Session-Id(263) l=57 f=-M- val=icscf.mnc001.mcc001.3gppnetwork.org;1013585100;16
  > AVP: Origin-Host(264) l=43 f=-M- val=icscf.mnc001.mcc001.3gppnetwork.org
  > AVP: Origin-Realm(296) l=37 f=-M- val=mnc001.mcc001.3gppnetwork.org
  > AVP: Destination-Realm(283) l=37 f=-M- val=mnc001.mcc001.3gppnetwork.org
  > AVP: Vendor-Specific-Application-Id(260) l=32 f=-M-
  > AVP: Auth-Session-State(277) l=12 f=-M- val=NO_STATE_MAINTAINED (1)
  > AVP: User-Name(1) l=41 f=-M- val=bob@mnc001.mcc001.3gppnetwork.org
  > AVP: Public-Identity(601) l=49 f=VM- vnd=TGPP val=sip:bob@mnc001.mcc001.3gppnetwork.org
    AVP Code: 601 Public-Identity
    > AVP Flags: 0xc0, Vendor-Specific: Set, Mandatory: Set
    AVP Length: 49
    AVP Vendor Id: 3GPP (10415)
    Public-Identity: sip:bob@mnc001.mcc001.3gppnetwork.org
    [SIP from address: sip:bob@mnc001.mcc001.3gppnetwork.org]
    Padding: 00000
  > AVP: Visited-Network-Identifier(600) l=41 f=VM- vnd=TGPP val=6d6e633030312e6d63633030312e336770706e6574776f72...
  
```

Fig. 2.20: UAR Diameter message (2) captured with protocol analysis software

```

Internet Protocol Version 4, Src: 10.0.2.150, Dst: 10.0.2.110
Transmission Control Protocol, Src Port: 49556, Dst Port: 3868, Seq: 1, Ack: 389, Len: 324
Diameter Protocol
  Version: 0x01
  Length: 324
  Flags: 0x40, Proxyable
    0... .... = Request: Not set
    .1.. .... = Proxyable: Set
    ..0. .... = Error: Not set
    ...0 .... = T(Potentially re-transmitted message): Not set
    .... 0.. = Reserved: Not set
    .... .0.. = Reserved: Not set
    .... ..0. = Reserved: Not set
    .... ...0 = Reserved: Not set
  Command Code: 300 User-Authorization
  ApplicationId: 3GPP Cx (16777216)
  Hop-by-Hop Identifier: 0x3d493cba
  End-to-End Identifier: 0x4ccdf17e
  [Request In: 17]
  [Response Time: 0.023716403 seconds]
  > AVP: Session-Id(263) l=57 f=-M- val=icscf.mnc001.mcc001.3gppnetwork.org;1013585100;16
  > AVP: Origin-Host(264) l=41 f=-M- val=hss.mnc001.mcc001.3gppnetwork.org
  > AVP: Origin-Realm(296) l=37 f=-M- val=mnc001.mcc001.3gppnetwork.org
  > AVP: Auth-Session-State(277) l=12 f=-M- val=NO_STATE_MAINTAINED (1)
  > AVP: Vendor-Specific-Application-Id(260) l=32 f=-M-
  > AVP: Server-Capabilities(603) l=84 f=VM- vnd=TGPP
    AVP Code: 603 Server-Capabilities
    > AVP Flags: 0xc0, Vendor-Specific: Set, Mandatory: Set
    AVP Length: 84
    AVP Vendor Id: 3GPP (10415)
  > Server-Capabilities: 0000025dc0000010000028af00000010000025ac0000038...
    > AVP: Optional-Capability(605) l=16 f=VM- vnd=TGPP val=1
    > AVP: Server-Name(602) l=56 f=VM- vnd=TGPP val=sip:scscf.mnc001.mcc001.3gppnetwork.org:5060
      AVP Code: 602 Server-Name
      > AVP Flags: 0xc0, Vendor-Specific: Set, Mandatory: Set
      AVP Length: 56
      AVP Vendor Id: 3GPP (10415)
      Server-Name: sip:scscf.mnc001.mcc001.3gppnetwork.org:5060
    > AVP: Experimental-Result(297) l=32 f=-M-

```

Fig. 2.21: UAA Diameter message (3) captured with protocol analysis software

```

Internet Protocol Version 4, Src: 10.0.2.120, Dst: 10.0.2.150
Transmission Control Protocol, Src Port: 3868, Dst Port: 36450, Seq: 1, Ack: 1, Len: 452
Diameter Protocol
  Version: 0x01
  Length: 452
  > Flags: 0xc0, Request, Proxyable
    Command Code: 303 Multimedia-Auth
    ApplicationId: 3GPP Cx (16777216)
    Hop-by-Hop Identifier: 0x6cbe32ff
    End-to-End Identifier: 0x4db7a58b
    [Answer In: 35]
    > AVP: Session-Id(263) l=56 f=-M- val=scscf.mnc001.mcc001.3gppnetwork.org;853087451;25
    > AVP: Origin-Host(264) l=43 f=-M- val=scscf.mnc001.mcc001.3gppnetwork.org
    > AVP: Origin-Realm(296) l=37 f=-M- val=mnc001.mcc001.3gppnetwork.org
    > AVP: Destination-Realm(283) l=37 f=-M- val=mnc001.mcc001.3gppnetwork.org
    > AVP: Vendor-Specific-Application-Id(260) l=32 f=-M-
    > AVP: Auth-Session-State(277) l=12 f=-M- val=NO_STATE_MAINTAINED (1)
    < AVP: Public-Identity(601) l=49 f=VM- vnd=TGPP val=sip:bob@mnc001.mcc001.3gppnetwork.org
      AVP Code: 601 Public-Identity
      > AVP Flags: 0xc0, Vendor-Specific: Set, Mandatory: Set
        AVP Length: 49
        AVP Vendor Id: 3GPP (10415)
        Public-Identity: sip:bob@mnc001.mcc001.3gppnetwork.org
        [SIP from address: sip:bob@mnc001.mcc001.3gppnetwork.org]
        Padding: 000000
    > AVP: User-Name(1) l=41 f=-M- val=bob@mnc001.mcc001.3gppnetwork.org
    > AVP: 3GPP-SIP-Number-Auth-Items(607) l=16 f=VM- vnd=TGPP val=1
    < AVP: 3GPP-SIP-Auth-Data-Item(612) l=40 f=VM- vnd=TGPP
      AVP Code: 612 3GPP-SIP-Auth-Data-Item
      > AVP Flags: 0xc0, Vendor-Specific: Set, Mandatory: Set
        AVP Length: 40
        AVP Vendor Id: 3GPP (10415)
        < 3GPP-SIP-Auth-Data-Item: 00000260c00001c000028af4469676573742d414b417631...
          > AVP: 3GPP-SIP-Authentication-Scheme(608) l=28 f=VM- vnd=TGPP val=Digest-AKAV1-MD5
    > AVP: Server-Name(602) l=56 f=VM- vnd=TGPP val=sip:scscf.mnc001.mcc001.3gppnetwork.org:5060
  
```

Fig. 2.22: MAR Diameter message (5) captured with protocol analysis software

```

Internet Protocol Version 4, Src: 10.0.2.150, Dst: 10.0.2.120
Transmission Control Protocol, Src Port: 36450, Dst Port: 3868, Seq: 1, Ack: 453, Len: 504
Diameter Protocol
  Version: 0x01
  Length: 504
  > Flags: 0x40, Proxyable
    Command Code: 303 Multimedia-Auth
    ApplicationId: 3GPP Cx (16777216)
    Hop-by-Hop Identifier: 0x6cbe32ff
    End-to-End Identifier: 0x4db7a58b
    [Request In: 27]
    [Response Time: 0.013702232 seconds]
    > AVP: Session-Id(263) l=56 f=-M- val=scscf.mnc001.mcc001.3gppnetwork.org;853087451;25
    > AVP: Origin-Host(264) l=41 f=-M- val=hss.mnc001.mcc001.3gppnetwork.org
    > AVP: Origin-Realm(296) l=37 f=-M- val=mnc001.mcc001.3gppnetwork.org
    > AVP: Auth-Session-State(277) l=12 f=-M- val=NO_STATE_MAINTAINED (1)
    > AVP: Vendor-Specific-Application-Id(260) l=32 f=-M-
    > AVP: Public-Identity(601) l=49 f=VM- vnd=TGPP val=sip:bob@mnc001.mcc001.3gppnetwork.org
    > AVP: 3GPP-SIP-Number-Auth-Items(607) l=16 f=VM- vnd=TGPP val=1
    > AVP: 3GPP-SIP-Auth-Data-Item(612) l=176 f=VM- vnd=TGPP
    > AVP: Result-Code(268) l=12 f=-M- val=DIAMETER_SUCCESS (2001)
  
```

Fig. 2.23: MAA Diameter message (6) captured with protocol analysis software

```

Internet Protocol Version 4, Src: 10.0.2.120, Dst: 10.0.2.150
Transmission Control Protocol, Src Port: 3868, Dst Port: 36450, Seq: 453, Ack: 505, Len: 476
Diameter Protocol
  Version: 0x01
  Length: 476
  > Flags: 0xc0, Request, Proxyable
    Command Code: 301 Server-Assignment
    ApplicationId: 3GPP Cx (16777216)
    Hop-by-Hop Identifier: 0x6cbe3300
    End-to-End Identifier: 0x4db7a58c
    [Answer In: 62]
  > AVP: Session-Id(263) l=56 f=-M- val=scscf.mnc001.mcc001.3gppnetwork.org;853087451;26
  > AVP: Origin-Host(264) l=43 f=-M- val=scscf.mnc001.mcc001.3gppnetwork.org
  > AVP: Origin-Realm(296) l=37 f=-M- val=mnc001.mcc001.3gppnetwork.org
  > AVP: Unknown(494) l=48 f=V-- vnd=50 val=37623963663964612d333336322d656561662d353336352d...
  > AVP: Destination-Realm(283) l=37 f=-M- val=mnc001.mcc001.3gppnetwork.org
  > AVP: Vendor-Specific-Application-Id(260) l=32 f=-M-
  > AVP: Auth-Session-State(277) l=12 f=-M- val=NO_STATE_MAINTAINED (1)
  > AVP: Public-Identity(601) l=49 f=VM- vnd=TGPP val=sip:bob@mnc001.mcc001.3gppnetwork.org
  > AVP: Server-Name(602) l=56 f=VM- vnd=TGPP val=sip:scscf.mnc001.mcc001.3gppnetwork.org:5060
  > AVP: User-Name(1) l=41 f=-M- val=bob@mnc001.mcc001.3gppnetwork.org
  > AVP: Server-Assignment-Type(614) l=16 f=VM- vnd=TGPP val=REGISTRATION (1)
    AVP Code: 614 Server-Assignment-Type
    > AVP Flags: 0xc0, Vendor-Specific: Set, Mandatory: Set
    AVP Length: 16
    AVP Vendor Id: 3GPP (10415)
    Server-Assignment-Type: REGISTRATION (1)
  < AVP: User-Data-Already-Available(624) l=16 f=VM- vnd=TGPP val=USER_DATA_NOT_AVAILABLE (0)
    AVP Code: 624 User-Data-Already-Available
    > AVP Flags: 0xc0, Vendor-Specific: Set, Mandatory: Set
    AVP Length: 16
    AVP Vendor Id: 3GPP (10415)
    User-Data-Already-Available: USER_DATA_NOT_AVAILABLE (0)

```

Fig. 2.24: SAR Diameter message (11) captured with protocol analysis software

```

Internet Protocol Version 4, Src: 10.0.2.150, Dst: 10.0.2.120
Transmission Control Protocol, Src Port: 36450, Dst Port: 3868, Seq: 505, Ack: 929, Len: 2444
Diameter Protocol
  Version: 0x01
  Length: 2444
  > Flags: 0x40, Proxyable
    Command Code: 301 Server-Assignment
    ApplicationId: 3GPP Cx (16777216)
    Hop-by-Hop Identifier: 0x6cbe3300
    End-to-End Identifier: 0x4db7a58c
    [Request In: 51]
    [Response Time: 0.054342403 seconds]
  > AVP: Session-Id(263) l=56 f=-M- val=scscf.mnc001.mcc001.3gppnetwork.org;853087451;26
  > AVP: Origin-Host(264) l=41 f=-M- val=hss.mnc001.mcc001.3gppnetwork.org
  > AVP: Origin-Realm(296) l=37 f=-M- val=mnc001.mcc001.3gppnetwork.org
  > AVP: Auth-Session-State(277) l=12 f=-M- val=NO_STATE_MAINTAINED (1)
  > AVP: Vendor-Specific-Application-Id(260) l=32 f=-M-
  > AVP: User-Name(1) l=41 f=-M- val=bob@mnc001.mcc001.3gppnetwork.org
  > AVP: Cx-User-Data(606) l=2144 f=VM- vnd=TGPP val=3c3f786d6c2076657273696f6e3d22312e302220656e636f...
    AVP Code: 606 Cx-User-Data
    > AVP Flags: 0xc0, Vendor-Specific: Set, Mandatory: Set
    AVP Length: 2144
    AVP Vendor Id: 3GPP (10415)
    Cx-User-Data: 3c3f786d6c2076657273696f6e3d22312e302220656e636f...
  > eXtensible Markup Language
    > <?xml
    > <IMSSubscription>
      > <PrivateID>
      > <ServiceProfile>
        > <PublicIdentity>
          > <InitialFilterCriteria>
            > <Priority>
            > <TriggerPoint>
          > <ApplicationServer>
            > <ServerName>
            > <DefaultHandling>
              </ApplicationServer>
            </InitialFilterCriteria>
          </ServiceProfile>
        </IMSSubscription>
  > AVP: Charging-Information(618) l=40 f=VM- vnd=TGPP
  > AVP: Result-Code(268) l=12 f=-M- val=DIAMETER_SUCCESS (2001)

```

Fig. 2.25: SAA Diameter message (12) captured with protocol analysis software

```

Internet Protocol Version 4, Src: 10.0.2.110, Dst: 10.0.2.150
Transmission Control Protocol, Src Port: 3868, Dst Port: 49556, Seq: 389, Ack: 325, Len: 300
Diameter Protocol
  Version: 0x01
  Length: 300
  > Flags: 0xc0, Request, Proxyable
    Command Code: 302 Location-Info
    ApplicationId: 3GPP Cx (16777216)
    Hop-by-Hop Identifier: 0x3d493ccb
    End-to-End Identifier: 0x4ccdf17f
    [Answer In: 121]
  > AVP: Session-Id(263) l=57 f=-M- val=icscf.mnc001.mcc001.3gppnetwork.org;1013585100;17
  > AVP: Origin-Host(264) l=43 f=-M- val=icscf.mnc001.mcc001.3gppnetwork.org
  > AVP: Origin-Realm(296) l=37 f=-M- val=mnc001.mcc001.3gppnetwork.org
  > AVP: Destination-Realm(283) l=37 f=-M- val=mnc001.mcc001.3gppnetwork.org
  > AVP: Vendor-Specific-Application-Id(260) l=32 f=-M-
  > AVP: Auth-Session-State(277) l=12 f=-M- val=NO_STATE_MAINTAINED (1)
  > AVP: Public-Identity(601) l=49 f=VM- vnd=TGPP val=sip:bob@mnc001.mcc001.3gppnetwork.org

```

Fig. 2.26: LIR Diameter message (15) captured with protocol analysis software

```

Internet Protocol Version 4, Src: 10.0.2.150, Dst: 10.0.2.110
Transmission Control Protocol, Src Port: 49556, Dst Port: 3868, Seq: 325, Ack: 689, Len: 276
Diameter Protocol
Version: 0x01
Length: 276
> Flags: 0x40, Proxyable
Command Code: 302 Location-Info
ApplicationId: 3GPP Cx (16777216)
Hop-by-Hop Identifier: 0x3d493ccb
End-to-End Identifier: 0x4ccdf17f
[Request In: 116]
[Response Time: 0.005661146 seconds]
> AVP: Session-Id(263) l=57 f=-M- val=icscf.mnc001.mcc001.3gppnetwork.org;1013585100;17
> AVP: Origin-Host(264) l=41 f=-M- val=hss.mnc001.mcc001.3gppnetwork.org
> AVP: Origin-Realm(296) l=37 f=-M- val=mnc001.mcc001.3gppnetwork.org
> AVP: Auth-Session-State(277) l=12 f=-M- val=NO_STATE_MAINTAINED (1)
> AVP: Vendor-Specific-Application-Id(260) l=32 f=-M-
> AVP: Server-Name(602) l=56 f=VM- vnd=TGPP val=sip:scscf.mnc001.mcc001.3gppnetwork.org:5060
    AVP Code: 602 Server-Name
    > AVP Flags: 0xc0, Vendor-Specific: Set, Mandatory: Set
    AVP Length: 56
    AVP Vendor Id: 3GPP (10415)
    Server-Name: sip:scscf.mnc001.mcc001.3gppnetwork.org:5060
> AVP: Result-Code(268) l=12 f=-M- val=DIAMETER_SUCCESS (2001)

```

Fig. 2.27: LIA Diameter message (16) captured with protocol analysis software

2.5 SAE (System Architecture Evolution) and LTE (Long Term Evolution)

As already outlined in Section 2.1 and especially in Table 2.1, a new, powerful radio access network, E-UTRAN (Evolved-UTRAN), was specified under the title LTE (Long Term Evolution) as part of the 3GPP Release 8. For the first time, it offered a purely packet-based air interface for all services, including the previously circuit-switched telephony. This provides the possibility of eliminating the circuit-switched core network. However, since the previous packet-switched GPRS core network was not designed for real-time capability, a new packet-switched core network, the EPC (Evolved Packet Core), had to be standardized for 3GPP Release 8 entitled SAE (System Architecture Evolution). In addition to the optional interface to the IMS, the EPC offers interfaces to the E-UTRAN, other IP-based access networks (Non-3GPP IP Access: e.g., WLAN, WiMAX (Worldwide Interoperability for Microwave Access), DSL), and also “normal” UTRAN. This makes it possible to use the access network technology that is particularly well suited or available. These revolutionary and yet evolutionary network changes are shown in Figure 2.28 [173].

On the left in Figure 2.28, the GSM (CS Domain) and GPRS core networks (PS Domain), as well as the GSM/GPRS (GERAN) and UMTS access networks (UTRAN) already described in Section 1.2, are shown as techniques that have been introduced some time ago. Concerning UMTS, they can be supplemented by IMS from Release 5 (see Table 2.1) and by WLAN access networks from Release 6 (see Table 2.1). From Release 8 (see Table 2.1), the network architecture and access network technology

were migrated, as shown in Figure 2.28 under the headlines “System Architecture Evolution (SAE)” and “Long Term Evolution (LTE)”. This means that a new core network technology for the PS Domain, the Evolved Packet Core (EPC), and a significantly more powerful Evolved-UTRAN access network technology have been standardized. The complete solution consisting of EPC and E-UTRAN is called Evolved Packet System (EPS) [173].

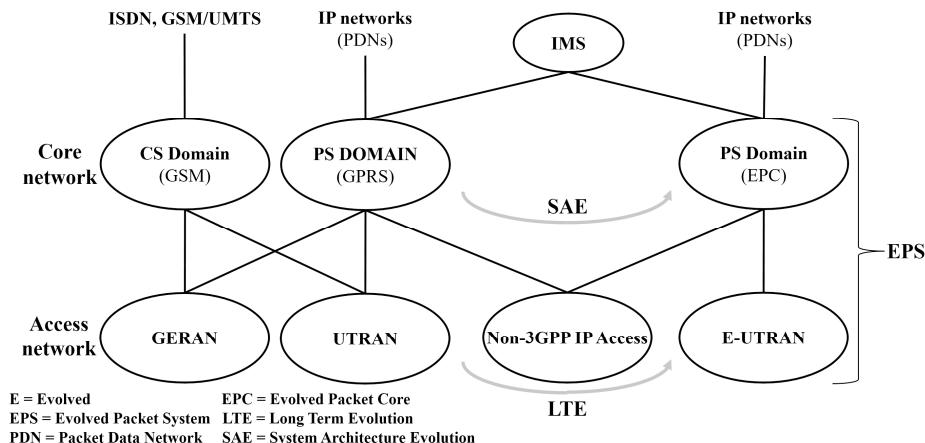


Fig. 2.28: Transformations in 3GPP networks towards SAE and LTE [173]

Advantageously, the various network technologies shown in Figure 2.28 can be used in combination in the sense of evolution, so that, for example, it is possible to gradually migrate from a GSM/GPRS to an All-IP network. Also, the various access networks can be applied side by side and in combination. This includes comprehensive mobility support, i.e., roaming and handover must be possible between the different access networks. Based on this, an uninterrupted, purely IMS-based service use is then possible despite a change of access network. Thus, from 3GPP Release 8 or the availability of the EPC, it is possible to make phone calls without interrupting the call despite switching between E-UTRAN, UTRAN, GERAN, and WLAN radio cells. With circuit-switched telephony, this could only have been achieved at a very high effort and expense. Besides, a pure IP network is cheaper in the long term anyway. So MSC/VLR and GMSC, as well as SGSN and GGSN, are still listed as core network technologies, but the associated GSM and GPRS networks are switched off in the medium or even short term [173].

While the radio access network (RAN) in the case of UTRAN with RNC and several connected NodeBs has a two-stage structure, in E-UTRAN, it is implemented in only one stage with the eNodeBs (eNBs), as shown in Figure 2.29. This reduces the

latency time, which is essential for real-time communication, and simplifies the network structure of the RAN, and makes it more flexible [173].

As Figure 2.29 also shows, the new IP core network, the EPC, to which the eNodeBs are connected, consists mainly of MME (Mobility Management Entity), S-GW (Serving-GW), PDN-GW (Packet Data Network). These network elements are supplemented by the PCRF (Policy and Charging Rules Function), which is also used in GPRS. Here as well, the NGN characteristic (see Section 1.3) of the separation of signaling and user data transport is consistently applied: the MME has signaling and control functions. The S-GW and PDN-GW provide all functions relating to user data.

The MME is responsible for:

- Complete signaling between UE and EPC for RAN-independent functions such as session and mobility management
- Security in the RAN
- Authentication of the UEs after querying the subscriber data in HSS
- Reachability of the UEs in idle state
- Assignment of the so-called EPS bearers, the user data channels
- QoS parameter negotiation
- Selection of S-GW and PDN-GW
- Roaming between access networks
- MME selection for handover.

The two gateways S-GW and PDN-GW in Figure 2.29 are responsible for IP user data transmission. The S-GW represents the router at the interface to E-UTRAN, supplemented by functions for lawful interception, QoS provisioning, and charging. Also, the S-GW acts as a mobile anchor point for handover between different eNodeBs or 3GPP access networks.

The PDN-GW represents the router at the interface to other IP networks and the IMS. It assigns IP addresses to the UEs, terminates the EPS bearers to the UEs, and provides subscriber-related firewall functionality as well as lawful interception, QoS provisioning, and charging. Besides, the PDN-GW also offers a mobile anchor point, but in contrast to the S-GW for mobility support between 3GPP and any other access network, such as WLAN.

The network element PCRF (Policy and Charging Rules Function) in Figure 2.29 provides policies for the user data streams and for charging. These rules are applied to the PDN-GW, i.e., data streams are rejected or allowed according to the policies of the PCRF and, in the latter case, also charged. To provide end-to-end QoS, the PCRF synchronizes QoS arrangements across network boundaries [173; 30].

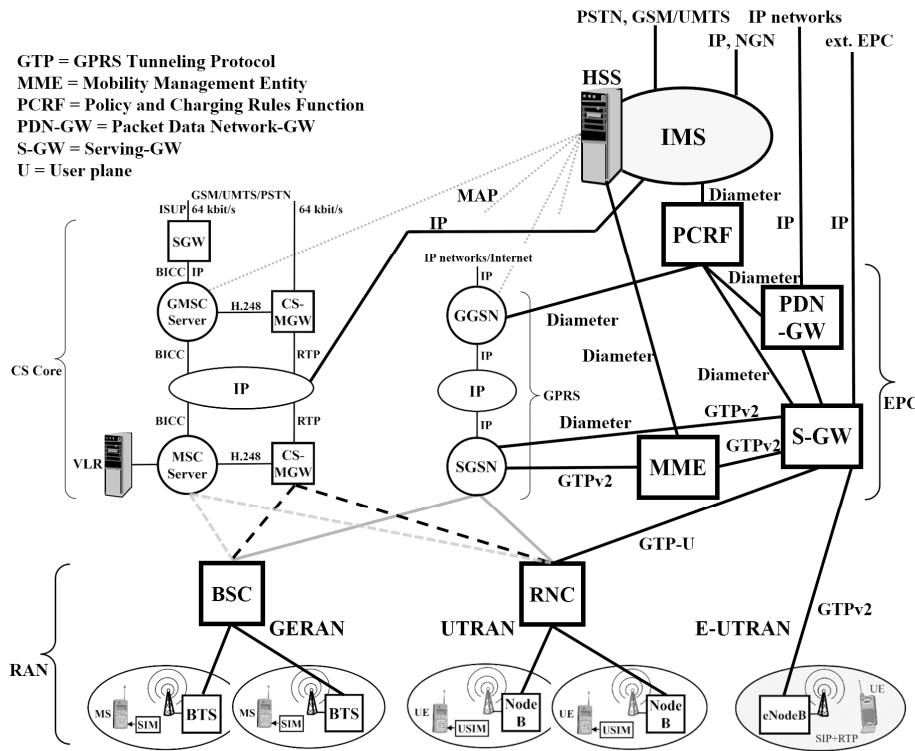


Fig. 2.29: 3GPP Release 8 mobile network

2.6 VoLTE (Voice over LTE)

As already explained in Section 2.5 regarding 3GPP Release 8, the LTE access networks and the EPC core network are IP-based only, i.e., they do not support voice communication per se. Therefore, telephony or real-time multimedia services in such an environment can be provided only with additional functionalities.

A first and technically very easy to realize possibility is the so-called OTT solution (Over The Top). Here the mobile network with E-UTRAN and EPC is only used for IP transport. The real-time services are provided on top with a separate system, possibly also by a different service provider (third party). This is a common method for delivering Voice over Internet but has the disadvantage that handovers are not supported at all, roaming is limited, and QoS cannot be guaranteed [173].

The second option, Circuit Switched Fall-back (CSFB), uses a GSM/UMTS infrastructure existing parallel to LTE and EPC with an additional GERAN or UTRAN connection of the mobile device (UE) and a CS domain (MSC) for telephony as shown in Figure 2.30. If an LTE user wants to make outgoing or incoming calls,

CSFB functionality transfers the connection to the 2G/3G network. For this purpose, the MSC must have an SGs interface to the EPC via upgrade. This enables registration of the UE via MME in the responsible MSC with combined access via LTE and UTRAN or GERAN. The UE requests a transfer to the CS network for an outgoing call. The MME then informs the eNodeB about the successful PS-CS transfer. When a call arrives via MSC, the MME notifies the UE (Paging). The procedure is then identical to that for outgoing calls. This solution, recommended for network operators without IMS and standardized in [35], requires - as outlined - extensions to UE, MME, MSC, and eNodeB. Nevertheless, it is relatively easy to implement. Handover and roaming are supported. The disadvantages are the mandatory 2G/3G network, significantly increased call setup times, and the lack of LTE data connectivity during a call [158; 173; 35].

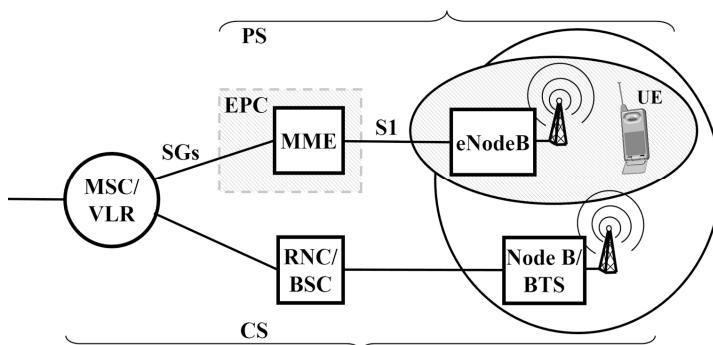


Fig. 2.30: Mobile network with LTE and Circuit Switched Fallback (CSFB) for telephony

The third option, “Voice over LTE over Generic Access Network (VoLGA)”, was specified by the VoLGA Industry Forum. This solution does not require any changes to both the access and core network. The functional enhancements as are necessary for telephony are provided by a gateway, the so-called VoLGA Access Network Controller (VANC), connected between the EPC and MSC. In addition to the unchanged 3GPP network elements, advantages include the simultaneous use of telephony and LTE-based data communication as well as SMS and emergency call support. Disadvantages are the additionally required gateway VANC and the missing standardization through 3GPP [173].

Despite the three options mentioned above for providing voice communication in an LTE network, the term “Voice over LTE (VoLTE)” applies specifically to the fourth option, the so-called “GSMA VoLTE Profile” (Groupe Speciale Mobile Association) or “IMS Profile for Voice and SMS” [114] with the first-time use of IMS for SIP-based telephony in LTE networks (see Section 2.2) [173].

[114] demands, among other things, the support of “Single Radio Voice Call Continuity (SRVCC)” in the 3GPP mobile network for uninterrupted handover when switching between PS E-UTRAN and CS UTRAN or GERAN. SRVCC was standardized by 3GPP [32; 34], is IMS-based, but also requires modifications in the UE and network enhancements, as shown in Figure 2.31.

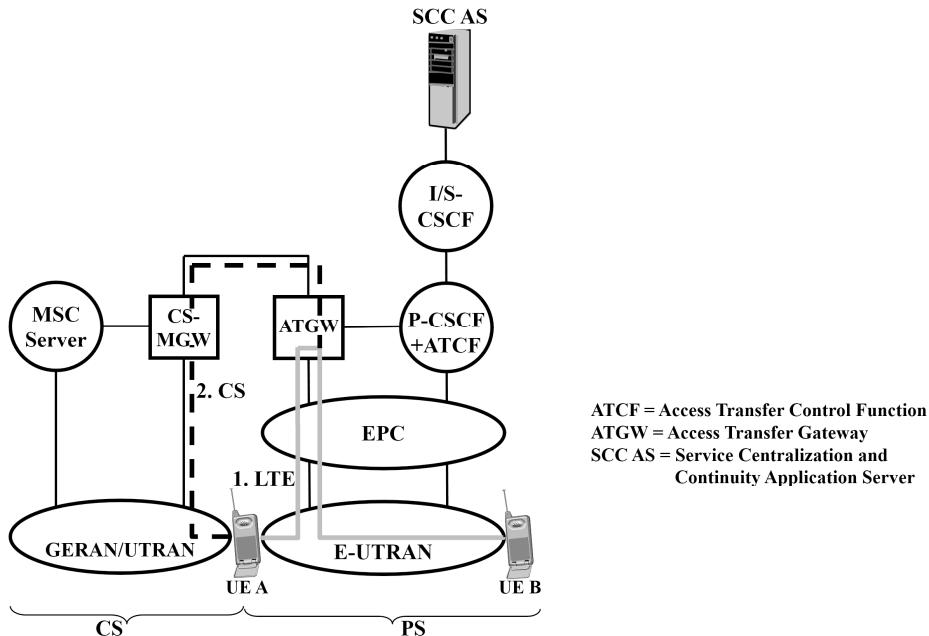


Fig. 2.31: VoLTE with SRVCC

For a possible transfer of a VoLTE call, the SCC AS (Service Centralization and Continuity Application Server) collects all information on an active SIP session. To ensure uninterrupted service in case of a necessary handover, the VoIP user data, the RTP streams, are not exchanged directly between the two participating UEs A and B, but via an ATGW (Access Transfer Gateway). This media gateway is controlled by an ATCF (Access Transfer Control Function), which is part of a P-CSCF, which, in turn, is connected to the SCC AS via S-CSCF. In the handover case of UE A, a CS-MGW converts the circuit-switched user data into RTP streams. The ATGW then switches between the original and the VoIP user data provided by conversion. A handover from the CS to the PS network is not supported [158; 34].

The advantages of this IP- and SIP-based VoLTE telephony solution with SRVCC in an EPS network are real multimedia over IP services, including roaming and

handover, when leaving the LTE radio cell coverage. The disadvantage is the high technical complexity [173].

Also, Voice over Wifi (VoWifi) is briefly discussed here as an example of non-3GPP IP access in Figure 2.28. Since VoLTE has no dependencies on the IP transport network in the EPC except the interface for QoS, it is relatively easy to integrate VoWifi. It can directly use the IMS. We only have to consider that the WLAN used to connect the UE usually is not trustworthy per se (Untrusted non-3GPP access). For this reason, a new network element, ePDG (Evolved Packet Data Gateway), has been introduced between the WLAN and the Internet, as shown in Figure 2.32. On the WLAN side, it implements a VPN gateway for IPsec tunnels to the UEs. On the EPC side, it provides the necessary MME and S-GW functions [158; 36].

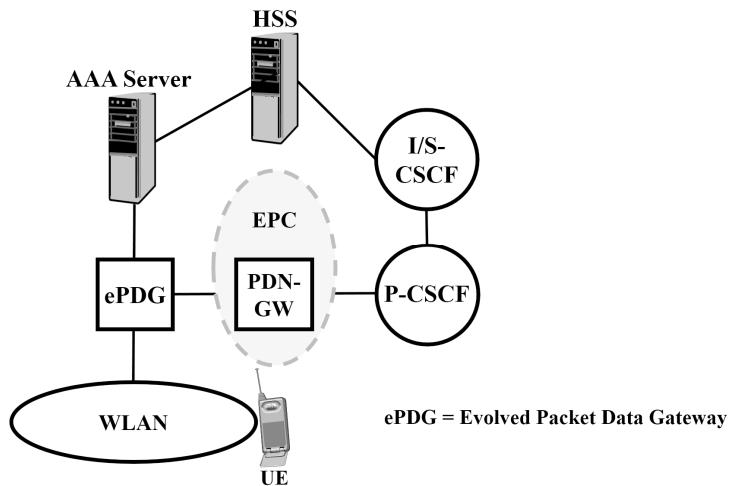


Fig. 2.32: VoWifi in 3GPP mobile network since Release 8

3 Future Networks

In the context of 4G, at the latest since 3GPP Release 14, the topic of virtualization entered the focus of mobile networks, i.e., the provision of network functions no longer in the form of explicit physical devices and systems but as virtual software functions based on standard computer hardware in, e.g., data centers. Since this requires flexibly manageable transport networks, SDN (Software Defined Networking) is more and more important. Also, parallel to the 3GPP standardization work for 4G mobile networks, the ITU-T has been working on a general concept for future networks under the keyword Future Networks. NFV (Network Functions Virtualization) and SDN play an essential role in this concept. Furthermore, the Future Networks concept seems to have been a key driver for 5G networks. Therefore, this chapter deals with NFV, SDN, and Future Networks, including their relationships in more detail.

3.1 NFV (Network Functions Virtualization) and MEC (Multi-access Edge Computing)

The functions of network elements or, more generally, network services such as firewalls or gateways are primarily implemented by software (SW). However, this software often still runs on special and, thus, proprietary hardware (HW), possibly based on a proprietary operating system (OS). Figure 3.1 shows this for a single network element such as the S-CSCF of an IMS (see Section 2.2) and for the network service CSCF, which combines the functionalities of a P-CSCF, I-CSCF, and S-CSCF. In the case of the CSCF, several interacting Network Functions (NF) provide the service. For this purpose, the individual services P-CSCF, etc., must be combined into one overall service by a central logic. This process is called orchestration (instrumentation). This concept, which was common in the past for the implementation of network elements and network services using proprietary hardware, results in comparatively high acquisition costs and relatively inflexible network architecture with mostly fixed functions [173].

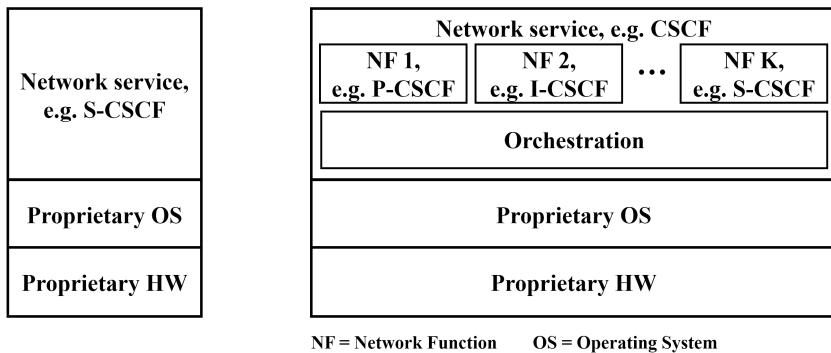


Fig. 3.1: Implementation of network elements or network services with proprietary hardware

The “Network Functions Virtualization (NFV)” concept was developed and standardized to overcome these disadvantages, especially from the point of view of network operators. It is based on the assumption that network functions are entirely implemented in SW and can, therefore, use standard hardware. As a result, proven IT virtualization techniques, such as the use of virtual machines (VM) and their joint operation on standard server hardware, can be applied [173].

The Industry Specification Group for NFV (ISG NFV) within ETSI addressed this issue in 2012 and defined NFV as follows: “Network Functions Virtualisation aims to transform the way that network operators architect networks by evolving standard IT virtualization technology to consolidate many network equipment types onto industry standard high volume servers, switches and storage, which could be located in Datacenters, network nodes, and the end user premises, ... It involves the implementation of network functions in software that can run on a range of industry-standard server hardware, and that can be moved to, or instantiated in, various locations in the network as required, without the need for installation of new equipment” [90].

Figure 3.2 illustrates these relationships and the resulting capabilities using a virtual IMS (see Section 2.2) with the NFs P-CSCF, I-CSCF, S-CSCF, HSS, etc.. SW instantiations of NFs run on virtual machines (VMs), whose number can be increased or decreased as needed. The VMs, in turn, use standard computer hardware abstracted via a virtualization layer realized, e.g., through a hypervisor. In addition to the illustration in Figure 3.2, VMs can also be implemented on separate hardware at different locations [173].

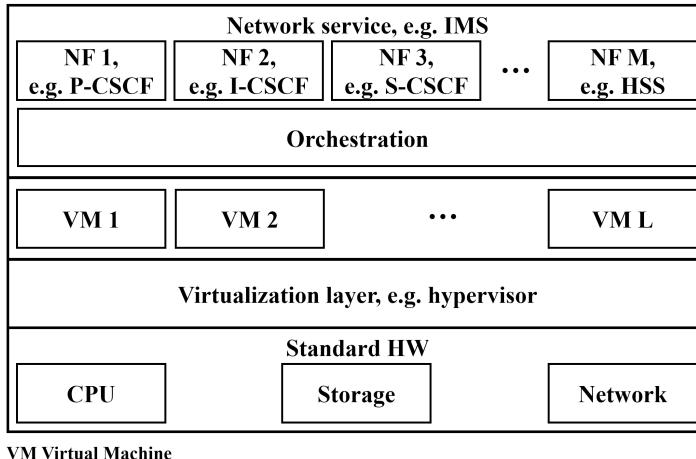


Fig. 3.2: NFV-based implementation of network elements or network services with standard HW using the example of IMS

The use of Network Functions Virtualization (NFV) can provide network operators with numerous advantages [90]:

- Lower equipment costs
- Faster introduction of new network capabilities and performance features, since only SW-, no longer HW-based
- Use of the same HW infrastructure for production, test, and reference environments
- High scalability
- Market access for software-only vendors
- Possibility to adapt the network configuration to the current traffic and its distribution in the network in real-time
- Use of the same HW by several network operators
- Lower electrical power consumption
- Lower planning, provision, and operating costs due to a homogeneous HW platform
- Automation of installation and operation by applying IT orchestration mechanisms and reusing VMs
- Simplification of the SW upgrade
- Gaining synergies between network operation and IT.

According to ETSI, Figure 3.3 [141] gives a first overview of the NFV framework. It consists of three areas, the Virtualized Network Functions (VNF) with the network services implemented in SW, the NFV Infrastructure (NFVI) for the virtualization of the VNFs based on physical hardware resources as well as the NFV Management

and Orchestration for the service composition from sub-services (orchestration) and the lifecycle management of the software, virtual and physical resources. A network service can be described with a single VNF or as a VNF set (e.g., to implement a pool of web servers without any relation between the VNFs) or as a so-called VNF Forwarding Graph (VNF-FG) to describe a network service formed by networking several VNFs (e.g., to access a web server via a firewall, NAT and load balancer). VNF-FG and VNF Set can also be combined. Furthermore, a VNF instance can run on different virtual and physical resources, even at various locations. A site with corresponding NFV resources is called NFVI-POP (NFV Infrastructure-Point of Presence). Usually, this is a data center [141].

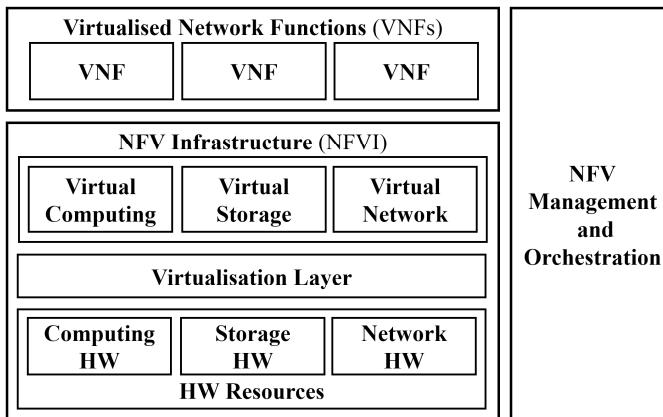


Fig. 3.3: Overview of the NFV framework [141]

As an extension of the overview in Figure 3.3, Figure 3.4 shows the complete NFV reference architecture framework, according to ETSI [141; 166]:

- NFVI (NFV Infrastructure): provides HW and SW resources for the VNFs. Virtualizes physical computing, storage, and networking resources
- VNF (Virtualized Network Function): virtual network function based on NFVI resources
- EMS (Element Management System): for configuration, analysis, and monitoring of a VNF
- NFV-MANO (NFV-Management and Orchestration): The NFV Orchestrator (NFVO) is responsible for the installation and configuration of new network services and the composition of the network services from VNFs. It receives the necessary information from the OSS/BSS (Operations/Business Support System) and, above all, from the “Service, VNF and Infrastructure Description”, which contains data on the VNFs (e.g., a VNF-FG), provisioning, and NFVI. This data is also used by the VNF Manager (VNFM) to manage the lifecycle of a VNF, i.e., in-

stantiation (creating a VNF), update/upgrade (new SW or changed configuration), required scaling (increasing or decreasing the capacity of a VNF, e.g., number of CPUs or VMs) and terminating (returning NFVI resources allocated by a VNF). Finally, the Virtualized Infrastructure Manager (VIM) is responsible for the allocation and management of virtual and physical resources, taking into account the interactions of a VNF with virtual computing, storage, and network resources. Performance, error, and capacity planning data are also captured.

- OSS/BSS (Operations Support System/Business Support System) [141; 132; 166].

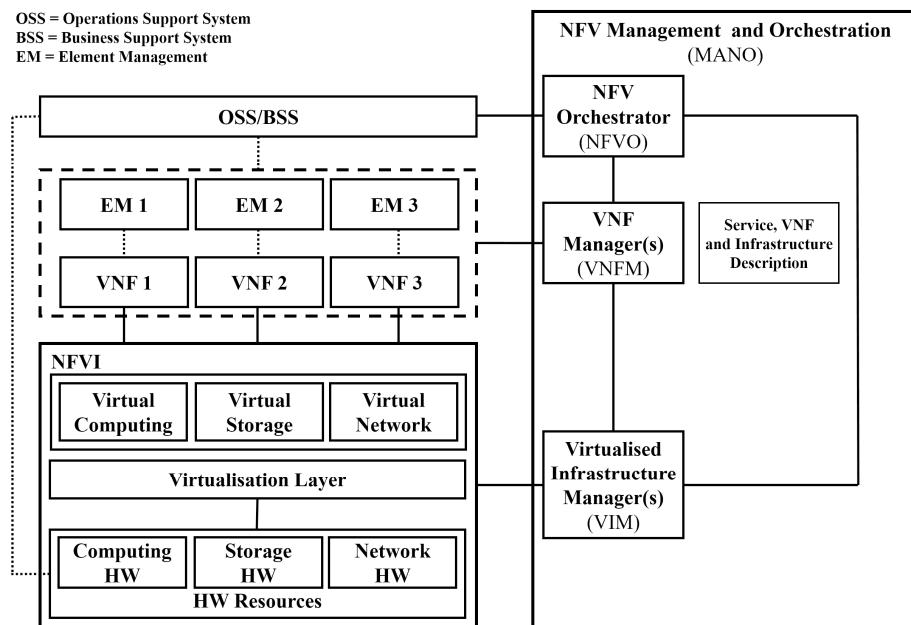


Fig. 3.4: NFV reference architecture framework according to ETSI [141; 132]

Besides, Figure 3.4 illustrates that the NFV Framework can be considered a 3-layer architecture: 1. NFVI + VIM, 2. VNFs/EMs + VNFM, 3. OSS/BSS + NFVO. This view allows a comparison with the cloud computing architecture of the ITU-T [182], according to [100] in Figure 3.5, which assumes four layers: 1. Resources and Network Layer, 2. Services Layer, 3. Access Layer, 4. User Layer. In this case, layer 1 is again divided into 1a. Physical Resources, 1b. Pooling and Virtualization, and 1c. Resource Orchestration. A direct comparison of the functionalities shows that Physical Resources + Pooling and Virtualization correspond to NFVI, Resource Orchestration to VIM(s) [166]. Also, one could argue about correspondences between IaaS (Infrastructure as a Service) or, above all, NaaS (Network as a Service) and the VNFs or

Service Orchestration and NFVO + VNFM. In this context, [140] speaks of a cloud service use case NFVIaaS (NFVI as a Service), the combination of IaaS and NaaS. This shows that there are strong similarities in functionality between NFV and cloud computing, and that cloud services can be provided based on an NFV framework, or that a cloud computing environment already provides many resources and functionalities for NFV.

However, even if the same technologies are used to a large extent in both approaches, the focus of NFV is clearly on the network. In cloud computing, the network is there to provide scalable IT services to a wide variety of customers – not primarily communications network operators and service providers. In this regard, there are significant differences, but the technologies used, such as flexible broadband IP networks, data centers, and, above all, virtualization has a high level of similarity [173].

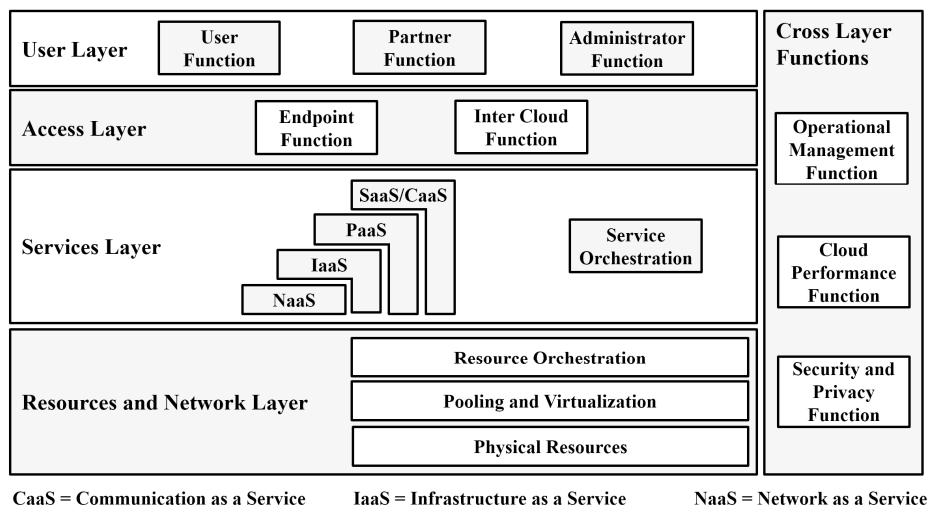


Fig. 3.5: Functional cloud computing reference architecture [100]

[140] mentions numerous fields of applications or use cases for NFV. We would like to point out in particular:

- The mobile IP core network EPC with MME, S-GW, and PDN-GW (see Section 2.5)
- The IMS with P/I/S-CSCF, HSS, PCRF (see Section 2.2)
- Mobile radio base stations like NodeB, eNodeB
- Content Delivery Networks (CDN) for the delivery of distributed and mirrored content, e.g., video streams.

[140] mentions the application of the VNF Forwarding Graphs (VNF-FG) explicitly. A VNF-FG describes a network service that is built by linking several VNFs (Service Chain), e.g., for accessing a web server via the service chain VNF1 Firewall – VNF2 NAPT Gateway – VNF3 Load Balancer. The VNFs are chained via logical connections (links). Figure 3.6 shows an example of two VNF-FGs with correspondingly two network forwarding paths (Service 1: VNF1 – VNF2 – VNF4 – VNF5 and Service 2: VNF1 – VNF2 – VNF3 – VNF4 – VNF5) for implementing two different network services based on NFVI [132; 141; 140]. Examples of this could be the use of the WWW service, i.e., the access to websites of WWW servers, by parents or children. In the first case, service chain 1 with VNF1 (Firewall) – VNF2 (Intrusion Detection System) - VNF4 (NAPT GW) – VNF5 (Load Balancer) is passed through. In the second case, with the children as users, service chain 2 with the additional VNF3 (parental control) is used.

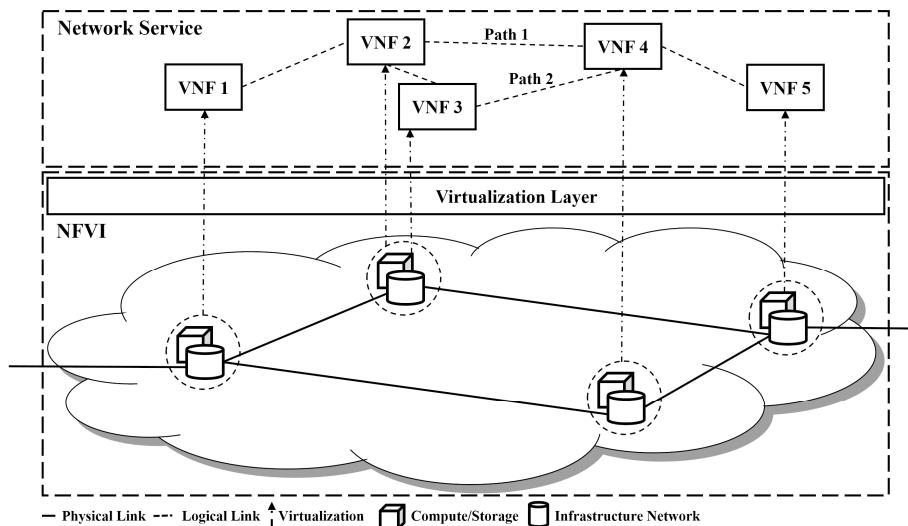


Fig. 3.6: Example of VNF-FG-based network services with VNFs in operation at different locations [132]

Of course, the chaining of network services (service chaining) to provide a communication service to a user is not new. This concept has already been used for many years in legacy networks, but using hard-wired HW-based network elements. The resulting functional chain must then always be passed through completely, even if this would not be necessary for some users. In the example mentioned above, this means that the Parental Control function would also have to be performed in the case of the parents.

NFV thus provides a virtualization solution for networks that need to offer high flexibility and dynamic changes in communication services by orchestrating the necessary network services. This includes the provision of the individual VNFs, but above all, their individual chaining. Virtual and physical resources can also be requested and allocated as needed. It is also possible to provide new network functions flexibly and dynamically, to increase, reduce or relocate the performance of existing network services, and to do this at different geographical locations (e.g., data centers). This results in corresponding requirements for the transport of messages between the VNFs, particularly concerning service chaining, since in this context, the destination of an IP packet, a network service, can or must change its IP address or, in the case of a service chain, intermediate destinations must be addressed in a particular order. It is, therefore, not sufficient to route an IP packet based on its IP destination address only. Switching and routing in an NFV environment is, therefore, the subject of a separate Section 3.2.

As already mentioned above, an important use case for NFV is the RAN, specifically the C-RAN (Cloud-RAN or also called Centralized-RAN). Here, as shown in Figure 3.7, the base station, e.g., the eNodeB for LTE, is divided into a BBU (Base Band Unit) and an RRH (Remote Radio Head, amplifier + RF filter + antenna(s)). It makes it possible to accommodate both at different locations, the RRH together with the antenna on the antenna mast and the BBU miles away at a more central location. This, in turn, means that several BBUs in a central cluster can be combined in a pool and provided on standard hardware using virtualization and, thus, NFV. Such a BBU pool then supplies many decentralized RRHs that are remotely connected via optical links (Fronthaul). The advantages of such a C-RAN solution are lower system, operating and upgrade costs, and energy savings [174; 140].

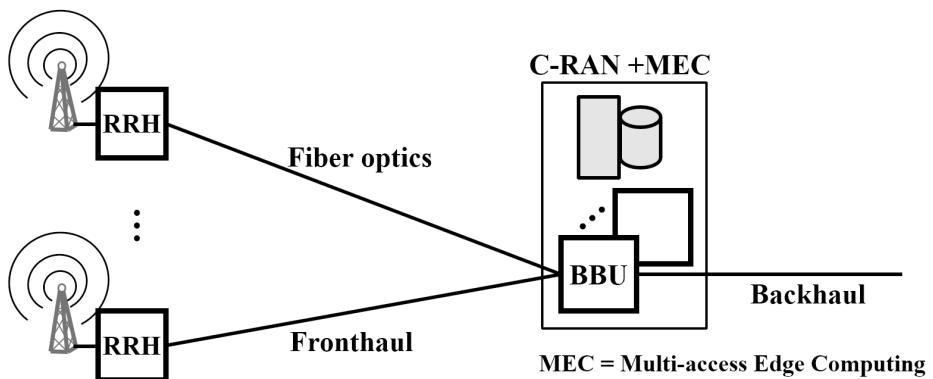


Fig. 3.7: C-RAN solution

While the C-RAN approach described above provides essential RAN functions in the cloud, Multi-access Edge Computing (MEC), standardized by ETSI ISG for MEC (Industry Specification Group) and described below, introduces cloud computing into the RAN, i.e., at the edge of the network. MEC thus offers service providers cloud computing functionalities close to the subscribers at or near the base stations to provide end users with applications in real-time, i.e., with very short delay times and high bit rates without involving the core network. This also implies that RAN operators can offer their computing resources to 3rd party providers for applications with corresponding requirements [174].

Applications for MEC include:

- Optimized video delivery
- Local content caching
- Augmented and Virtual Reality
- Gaming
- Car-to-x communication
- IoT gateway [136; 138].

ETSI standardized the framework required for MEC; Figure 3.8 gives an overview. A so-called MEC host provides decentralized cloud computing functionalities. It offers a virtualization infrastructure comparable to the NFVI in Figure 3.4, on which the MEC-SW applications are executed. These, in turn, can use special MEC services provided by the MEC platform. This includes information about the conditions at the radio interface, the location, or the allocation of a dedicated bit rate. Management and orchestration of the application SW, as well as the virtual and physical resources, are handled by MEC Management, if necessary, from a more central location [137; 95].

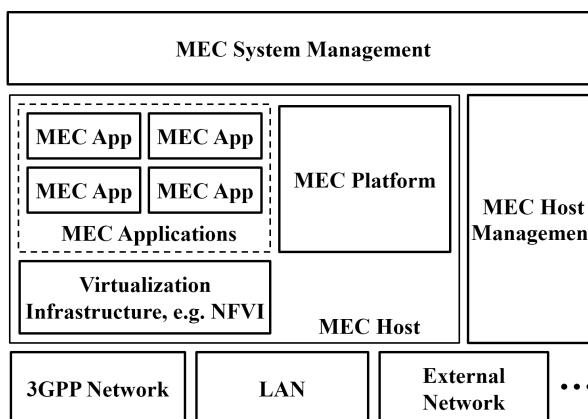


Fig. 3.8: Multi-access Edge Computing framework [137]

Figure 3.9 shows the MEC reference architecture in more detail, according to [137]. Here it can be seen that MEC management is divided into three levels: Virtualization Infrastructure Manager (VIM), MEC Platform Manager, and MEC Orchestrator. These are further significant parallels to the NFV reference architecture in Figure 3.4. In addition, [137] specifies a reference architecture for a MEC solution integrated into NFV. If this is the case, NFVO, VNFM, and VIM of the NFV environment control the decentralized MEC frameworks available in the network. Therefore, NFV and MEC complement each other and use the same techniques.

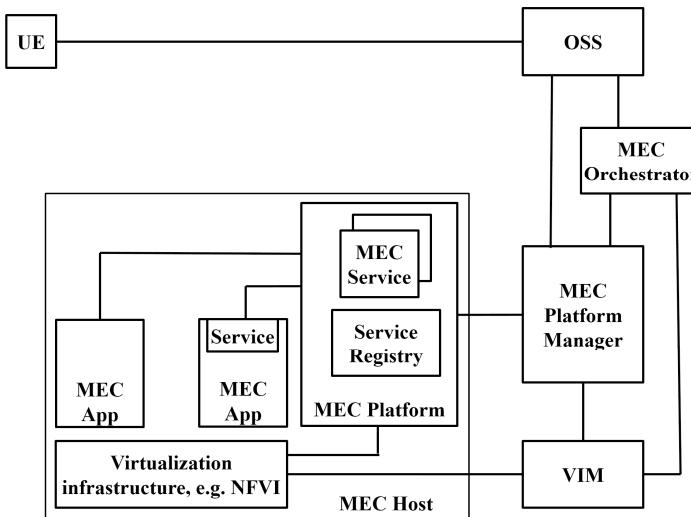


Fig. 3.9: MEC reference architecture [137]

3.2 SDN (Software Defined Networking) and SFC (Service Function Chaining)

The dynamic instantiation and migration of network functions in the framework of NFV also generate new challenges for Ethernet and IP transport networks. Depending on the network situation (e.g., peak traffic loads), data packets or data flows from dynamically relocated and/or newly scaled network applications must be flexibly forwarded to the responsible network services (e.g., IMS-VNFs) in the NFV infrastructure (e.g., to NFVI-POPs in various data centers) [173].

If a specific VNF, e.g., a DNS server, is migrated from HW server 1 to HW server 2 within the LAN of a data center A, or even migrated from an IP subnet in data center A to a remote data center B, this server must still be accessible by the end systems using it via the original MAC or IP address. This requires the application of tunnel-

ing procedures, i.e., overlay networks are formed in the existing transport networks, in which the original addresses can still be used. Usual mechanisms for this are:

- VLAN (Virtual LAN)
- VXLAN (Virtual Extensible LAN)
- GRE (Generic Routing Encapsulation)
- MPLS (Multiprotocol Label Switching).

With VLAN [69], virtual, i.e., logical LANs, are established based on a physical LAN. If there is a change in the NFV infrastructure, only the associated VLAN needs to be adapted accordingly. The VLAN ID or VLAN tag is changed, the MAC addresses remain unchanged. VXLAN offers a more scalable solution [14]. Here the L2 Ethernet frames with VXLAN IDs are encapsulated in UDP datagrams to form so-called VXLAN tunnels. The L2 traffic can then be easily tunneled through L3 WANs (e.g., between data centers) in UDP/IP packets. From the perspective of NFV infrastructure, virtual L2 overlay networks are used. GRE [1; 16] follows a generalized approach where any protocol can be encapsulated with a GRE header. This enables IP or Ethernet tunneling with IP or Ethernet over GRE over IP. MPLS [2] operates at the interface between L2 and L3. Each IP packet is labeled at the transition from an IP to an MPLS network, i.e., an additional label is applied, whereby packets that belong to the same flow get the same label. As a result, the complete IP headers no longer have to be evaluated but only the MPLS labels. Layer 2 forwarding is used instead of layer 3 routing. All data packets with the same label take the same route through the network. Labels are only valid in segments; in the beginning, they must be assigned, i.e., distributed. Due to the label use and its validity only in sections, VPNs (Virtual Private Networks) can be realized very easily. They can be used as flexibly adaptable overlay networks in an NFV environment [173].

The situation becomes more complicated if – as described in Section 3.1 – service chains are to be provided because, in these cases, intermediate destinations must be accessed according to the sequence of the network services as specified by the VNF-FG. It is, therefore, not sufficient to route an IP packet based on its IP destination address only. Two possible solutions to this problem are

- SDN (Software Defined Networking) and
- SFC (Service Function Chaining).

SDN is considered a key technology for the required, flexible handling of data flows in the context of NFV. While NFV decouples software and hardware of the network services, SDN separates the control logic with the associated signaling (control plane) from the user data (user plane or data plane) in the network nodes of the IP transport network, i.e., in switches and routers, as shown in Figure 3.10. This is done by introducing a central SDN controller for the control plane and decentralized, pure SDN switches reduced to data packet forwarding [173].

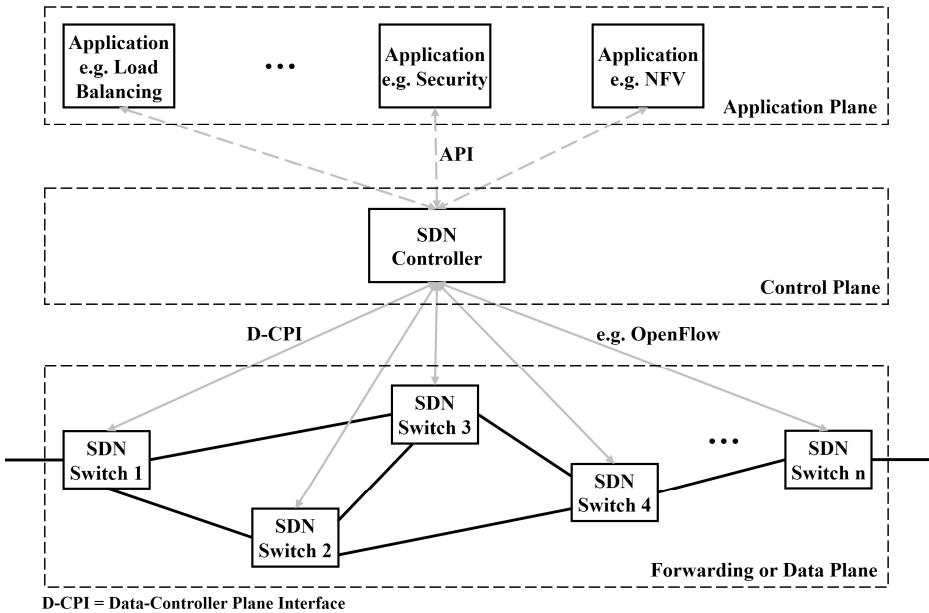


Fig. 3.10: SDN architecture

The so-far monolithic switches and routers are divided by a layer structure into simple SDN switches (in the data plane), which are only responsible for the forwarding of data packets, and SDN controllers (in the control plane), which provide the control logic (separation of control plane and data plane). Concerning flexibility and costs, an SDN controller usually controls a whole range of SDN switches via the Data-Controller Plane Interface (D-CPI), e.g., using the OpenFlow protocol. The protocol processing, i.e., the decision of what to do with a flow, a sequence of related data packets, takes place in the SDN controller (central logical control, can still be distributed over several physical or virtual, even redundant network nodes). The rules for forwarding are transmitted to the SDN switches by the SDN controller, e.g., via OpenFlow. Such simple SDN switches no longer need to be able to understand and evaluate numerous protocols; they must mainly support communication with the SDN controller in addition to packet forwarding. This reduces costs; the dependence on specific vendors is reduced. Besides, a complete transport network controlled by an SDN controller, consisting of many SDN switches, is logically a single switch or router and, therefore, easier to administer. Furthermore, Figure 3.10 shows another advantage of the SDN concept. Via APIs (Application Programming Interface), SDN applications (in the application layer) can program the SDN controller to change its behavior at runtime and thus implement new network services in the short-term (programmability). The SDN concept enables network administrators to configure and manage the transport network flexibly and dynamically from a cen-

tral point. This ensures security and optimized use of network resources utilizing self-developed SDN programs. SDN applications can be used for switching, routing, load balancing, QoS, traffic engineering, security, NFV, etc. Provisioning and orchestration systems (see Section 3.1) can also be connected via these APIs [173; 163; 89].

The use of SDN with the four characteristics “separation of control plane and data plane”, “central logical control”, “open interfaces”, and “programmability” can provide network operators with numerous advantages:

- Centralized, simultaneous, and consistent control of all switches
- Use of switches from various vendors
- Central overall view of the network
- Orchestration and management tools for automated and rapid deployment, configuration, and system updates across the network
- Programming of the network in real-time
- Improved network reliability and security
- Fine-grained handling of different data flows
- Easier adaptation of the network to user requirements [173].

In the following, we will examine the interrelationships of SDN in more detail, whereby OpenFlow as the control protocol is assumed.

The rules of how an SDN switch should handle different flows, i.e., data packets that belong together (with, e.g., the same IP source and destination addresses), are stored in the so-called Flow Table. Based on this, SDN works as follows:

1. The SDN Controller configures the SDN Switch with flow table entries.
2. The SDN switch analyses received data packets and checks them for matches with the flow table entries. If there is a match, the intended action, e.g., the requested forwarding, is executed.
3. If there is no match, the SDN switch forwards the packet via, e.g., OpenFlow protocol to the SDN controller to determine the processing.
4. Subsequently, the SDN controller will update the flow table in the SDN switch with a new entry so that the previously unknown data flow can now be processed locally in the SDN switch. Wild cards can also be set for a whole range of different data flows [102].

Figure 3.11 shows, on the one hand, the internal structure of an SDN switch supporting OpenFlow and, on the other hand, an SDN architecture in which not only one but several SDN controllers (here 2) work together with the same OpenFlow switch. This illustrates that a transport network based on SDN can consist of N switches and M controllers. The interface between a switch and a controller is called the OpenFlow channel. In addition to this interface for control, an SDN switch provides interfaces or ports to other switches, networks, and/or connected systems, or end devices such as computers in a data center. Based on a physical port, logical ports can also be created, e.g., for link aggregation, tunneling, or loopback. The

Ethernet or IP packets received and sent via the ports are further processed according to the contents of the tables, also shown in Figure 3.10, the so-called Flow and Group tables. As mentioned above, a received packet is analyzed whether it belongs to an already known flow with the same identifying parameters like MAC source and destination address and/or IP source and destination address and/or VLAN ID, port source and destination number, TCP or UDP-L4 protocol, etc. If this is the case, it is handled according to the rules defined in the flow table for this flow. E.g., in the case of switching or routing, it is forwarded through the network via the corresponding switch port on its path. For more complex evaluations or to take various parameters into account, several flow tables can also be passed through one after the other in a pipeline. If several flows are affected by the same actions, these flows, which represent a group of flows, are forwarded to a group table and processed there according to the defined actions. Also, the OpenFlow architecture includes so-called Meter Tables, which specify and implement metrics to be logged and adhered to per-flow, such as a permissible peak bit rate, thus enabling the implementation of even more complex QoS operations. In this case, a flow table could refer to a meter table, which in turn could refer to a flow table [149; 166].

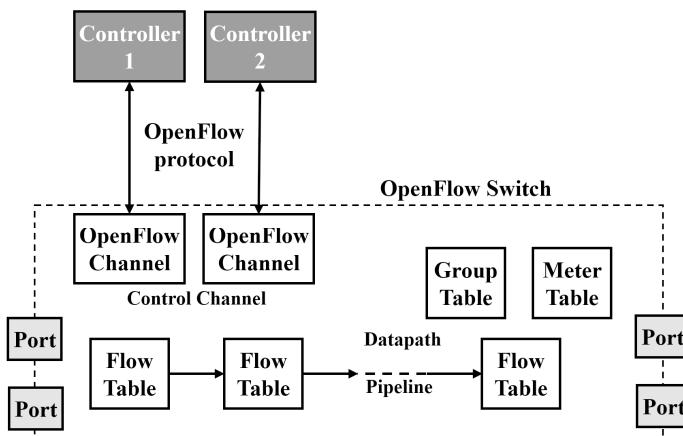


Fig. 3.11: OpenFlow switch [149]

The functionality of a flow table can be described even more concretely using Figure 3.12. Consequently, it consists of 7 fields with specific flow table entries [149; 166]:

- Match Fields: An incoming data packet is checked for a match with the values defined here.
- Priority: Indicates the priority of this flow table
- Counters: This counter is always incremented when an analyzed data packet meets the values in match fields.

- Instructions: Specifies the actions to be applied to the data packet if a match is detected, or forwards the packet to a subsequent flow table during pipelining
- Timeouts: Specifies the maximum idle time for flow before it is declared as no longer existing
- Cookie: Value selected by the controller, which can be used to filter flows for statistics, flow modifications, or deletions
- Flags: This can be used to modify flow table entries, e.g., a flow can be deleted by a corresponding flag.

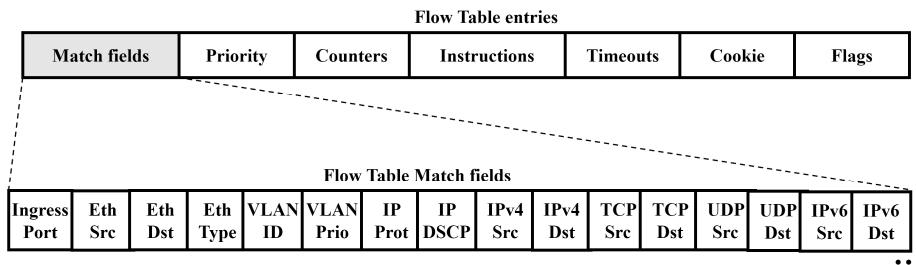


Fig. 3.12: Flow table [166]

The match fields can contain, among others, the fields shown in Figure 3.12. A received data packet is checked for their content [149; 166]: Switch Input Port, Ethernet Source Address (MAC address), Ethernet Destination Address, Ethernet Type Field (protocol transported in the IP packet (e.g., TCP)), VLAN ID, VLAN Priority, IP protocol (IPv4 or IPv6), IP DSCP (Differentiated Services Code Point, to determine the priority of an IP packet concerning QoS), IPv4 Source Address, IPv4 Destination Address, TCP Source Port, TCP Destination Port, UDP Source Port, UDP Destination Port, IPv6 Source Address, IPv6 Destination Address, etc.

A set of instructions is assigned to each flow table entry. Instructions describe the OpenFlow processing that happens when a packet matches the flow entry. An instruction either modifies pipeline processing, such as directing the packet to another flow table or contains a set of actions to add to the action set, or contains a list of actions to apply immediately to the packet.

An action is an operation that acts on a packet. An action may forward the packet to a port, modify the packet (such as decrementing the TTL (Time To Live) field) or change its state (such as associating it with a queue). Most actions include parameters; for example, a set-field action includes a field type and field value.

The instruction types are [149; 72]:

- Apply-Actions (optional): apply a list of actions to a packet immediately
- Clear-Actions: clears all the actions in the action set immediately
- Write-Actions: merges the specified set of action(s) into the current action set

- Write-Metadata (optional): writes the masked metadata value into the metadata field
- Stat-Trigger (optional): generate an event to the controller if some of the flow statistics cross one of the threshold values
- Goto-Table: indicates the next table in the processing pipeline.

A list of actions is an ordered list of actions included in a flow entry in the Apply-Actions instruction or a packet-out message, and that is executed immediately in the listed order, e.g., change IP destination address, then change MAC destination address, then send the packet to port X, then again modify destination IP address and send the packet via port Y [72]. An action set, empty by default, represents a set of actions associated with the packet that is accumulated while the packet is processed by each table and that is executed in the standard specified order when the instruction set terminates pipeline processing.

The OpenFlow standard specifies the following actions [149; 166]:

- Output: Forwarding a packet to a defined port, to another switch, or to a controller
- Group: Further processing of the packet according to a group table
- Drop: Stop handling the packet, drop it
- Set-Queue (optional): Specifies the queue to be used to send the packet to a port. To support QoS
- Meter (optional): Forward the packet to a meter table
- Push-Tag/Pop-Tag (optional): Adding or removing a tag, e.g., a VLAN ID or an MPLS label
- Set-Field (optional): Overwriting values in the protocol header fields of a packet
- Copy-Field (optional): Copying values of header fields during pipelining
- Change-TTL (optional): Modifying the TTL (Time To Live) value for IPv4 and MPLS or the Hop Limit value for IPv6.

As already mentioned, an SDN switch can contain not only one but several flow tables, which are then processed one after the other in a defined order in a pipeline. According to Figure 3.13, a distinction is made between two phases of processing: Ingress processing, starting from the inbound port through which a packet was received, and egress processing, which takes place when the outbound port is determined. Ingress processing always takes place; egress processing is optional [149].

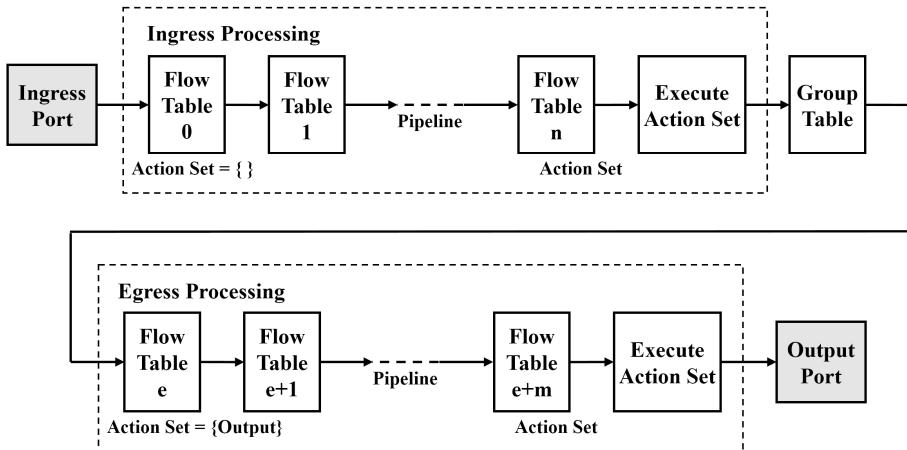


Fig. 3.13: Packet processing in the pipeline [149]

Group tables with the structure shown in Figure 3.14 can also be involved in the processing. It provides so-called action buckets for a whole group of flows, even considering several ports, whereby one action set per action bucket is stored for execution [149].

Group Identifier	Group Type	Counters	Action Buckets
------------------	------------	----------	----------------

Fig. 3.14: Group table [149]

Figure 3.15 shows the processing of a packet received by an OpenFlow Switch.

First, the ingress processing starts with the flow table 0. If one or more matches are detected, the corresponding actions are executed or stored in the action set. If necessary, the system proceeds to the next ingress flow table. If this was the last ingress flow table in the process flow, the corresponding action set is executed before the packet is forwarded, and, if necessary, the action set of the following group table is also executed. If there is also egress processing, it follows a similar procedure as for inbound processing, except the group table actions. For packets without matches, there must also be a flow table entry on how to handle such packets, e.g., that they are forwarded to a controller. If there are no instructions for this case, the packet is dropped, as shown in Figure 3.15 [149; 166].

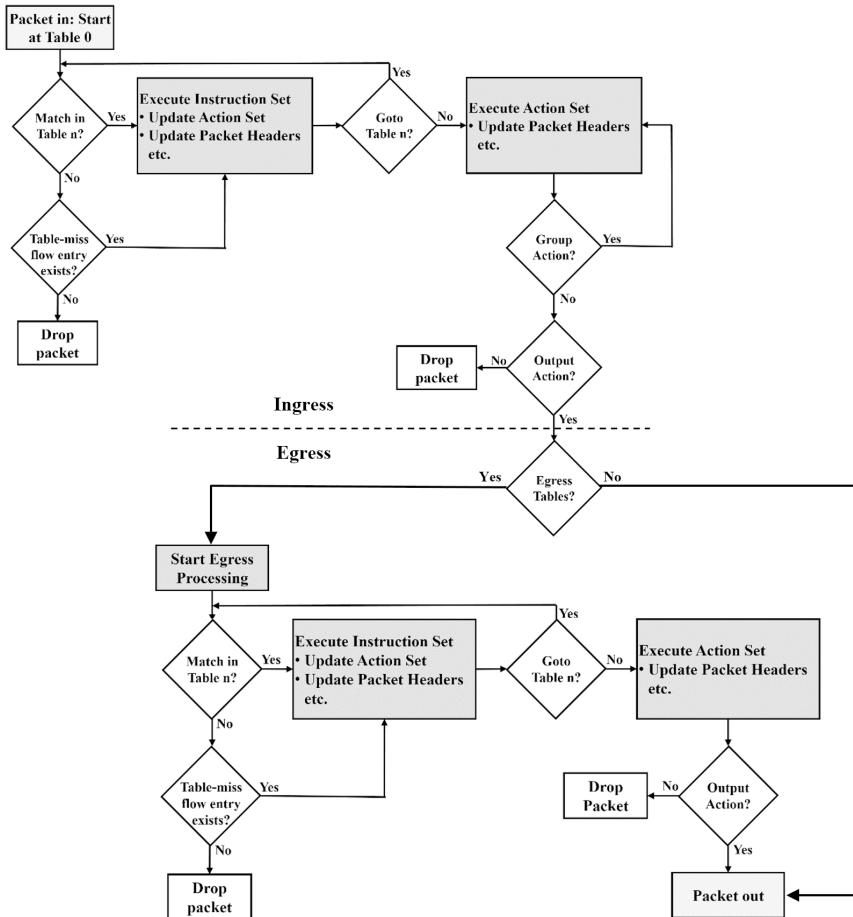


Fig. 3.15: Packet processing in OpenFlow switch [149]

Figure 3.16 illustrates the packet processing in three OpenFlow switches using a simple, practical example with IP routing based on SDN. If any values are allowed in the flow tables, they are marked with an *.

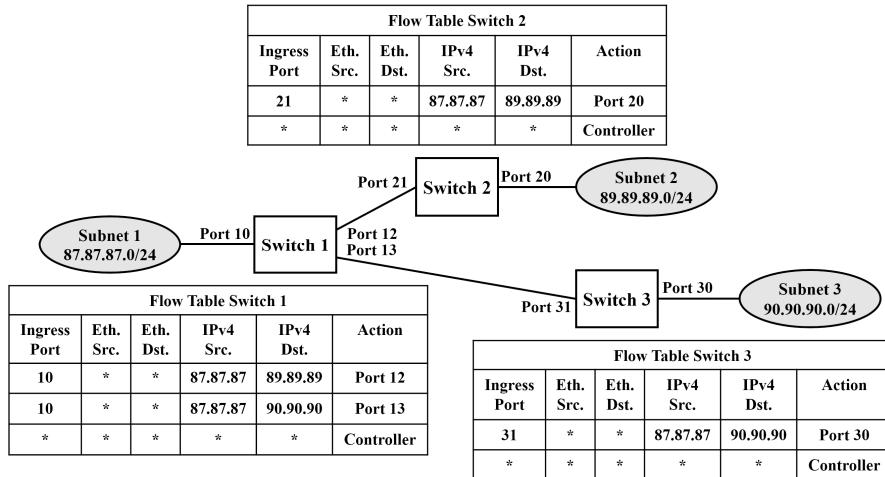


Fig. 3.16: Example for IP routing with SDN switches

According to Figure 3.17 [102], the communication between the SDN switch and SDN controller takes place via the D-CPI (Data-Controller Plane Interface, Southbound API), the so-called secure channel, through OpenFlow messages that are transmitted encrypted and connection-oriented based on TLS (Transport Layer Security) and TCP. Among the numerous specified OpenFlow messages, a distinction is made between the types controller-to-switch, asynchronous, and symmetrical. Controller-to-Switch messages are initiated by the SDN controller to configure the switch, query its capabilities, and manage the flow table. Asynchronous OpenFlow messages are initiated by the SDN switch to transmit a packet to the controller for which there is no flow table entry or to inform about status changes or errors. Finally, symmetrical messages can originate from both sides. This is used to establish an OpenFlow connection or to report an error [149; 173]. Table 3.1 [149; 166] shows all OpenFlow messages and briefly describes their functions.

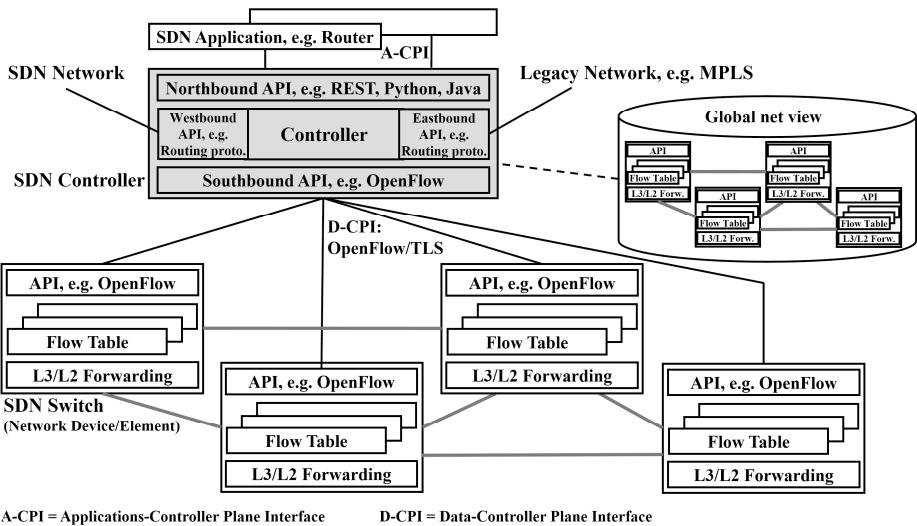


Fig. 3.17: SDN controllers and SDN switches in the network

Tab. 3.1: OpenFlow messages [149; 166]

Message	Function
Controller-to-Switch	
Features	The controller queries the request capabilities of a switch. The SDN switch responds with Features Reply, communicating its identity and supported capabilities.
Configuration	Configuration parameters in the switch are set or queried by the controller. Switch only responds to a query.
Modify-State	Entries in a flow or group table are added, deleted, or modified by the controller. The properties of the switch ports can also be set.
Read-State	For querying the current configuration, as well as statistical and performance data of a switch by the controller
Packet-out	The controller uses it to send data packets via its defined switch port, e.g., after analysis of the packet received with Packet-in.
Barrier	The controller determines with a Barrier request message confirmed by the switch with a reply whether a specific OpenFlow process was successfully completed.
Role-Request	For setting or querying the OpenFlow channel or the controller ID in a switch by a controller, especially when using several controllers
Asynchronous Configuration	For setting or querying a message filter by the controller for the own OpenFlow Channel, especially when using several controllers in the same transport network

Message	Function
Asynchronous	
Packet-in	The switch sends a data packet to the controller.
Flow-Removed	Switch informs the controller that a flow table entry has been removed.
Port-Status	Switch informs the controller about changes in the configuration of a port.
Role-Status	The switch indicates the controller change of role, e.g., that this controller is no longer the master controller.
Controller-Status	Switch informs the controller about changes in the status of the OpenFlow channel.
Flow-monitor	Switch informs the controller about change in a flow table.
Symmetrical	
Hello	Are exchanged when the connection between controller and switch is established
Echo	With the Echo request and the resulting reply message, the active connection is indicated. Can be initiated by switch or controller
Error	Error indication by controller or switch
Experimenter	For experimenting with functions not yet standardized

Figure 3.18 shows an example of an OpenFlow session. In the first step, a TCP connection between the SDN switch and the controller is established. Based on this, a secure channel is provided by TLS. The OpenFlow messages are then exchanged in this secure channel. It starts with Hello messages to establish an active OpenFlow connection. Subsequently, the SDN controller queries the capabilities supported by the switch with Feature Request. It responds with the desired information in a Feature Reply message. In the next step, the SDN controller makes new entries in the flow table of the switch with Modify-State. The same message but with different parameters can also be used to modify or delete entries in the flow table. If the switch has no entry in its flow table for a received packet, it forwards it with the message Packet-in to the controller for evaluation, which returns it with Packet-out to the switch after processing. In most cases, this is the first packet of a new flow to be handled accordingly in the future. In this case, the controller changes the corresponding entry in the flow table with Modify-State. The exchange of the OpenFlow messages Echo request and Echo reply indicates an active connection [102; 163; 173].

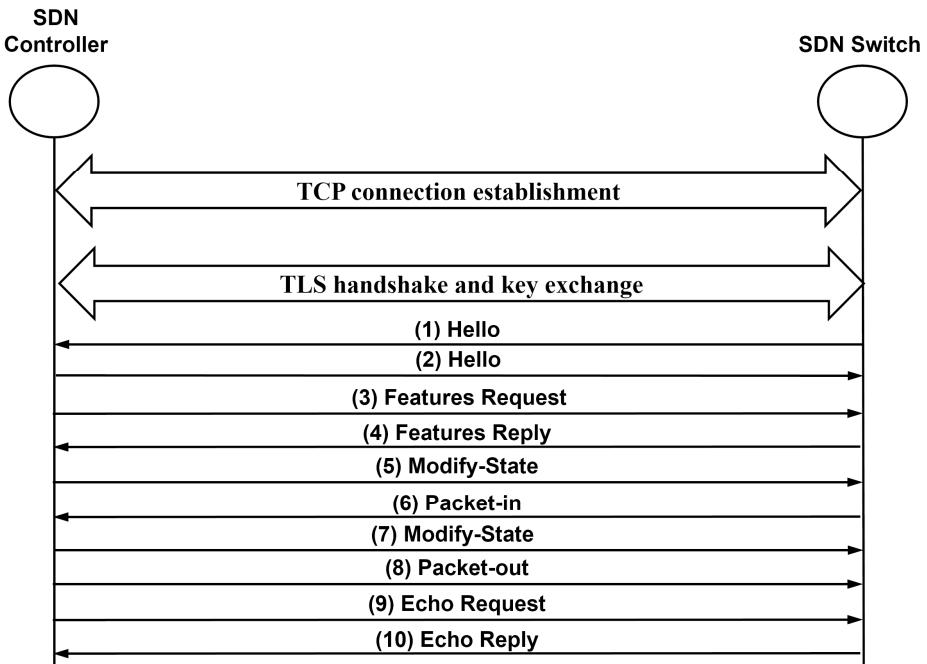
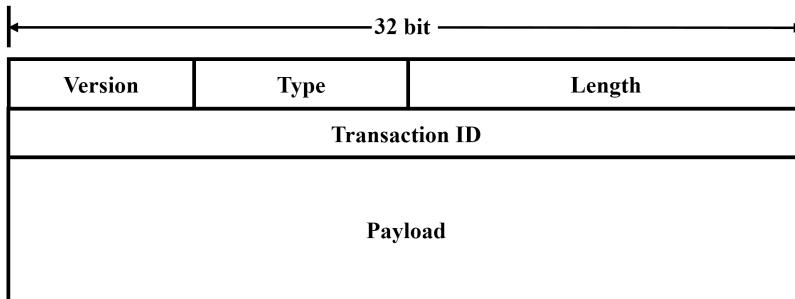


Fig. 3.18: Exemplary OpenFlow message exchange

In addition to the OpenFlow message sequence chart in Figure 3.18, the following Figures 3.19 to 3.23 describe the structure of OpenFlow messages in more detail. Figure 3.19 shows the basic structure with header and payload, arranged in 32-bit lines. The first two lines describe the header, which is the same for all OpenFlow messages. It contains fields for the protocol version (8 bit, e.g., 1.4), Type to identify the OpenFlow message type used (8 bit, e.g., OFPT_HELLO for Hello), Length (16 bit) for the length specification of the entire OpenFlow message incl. the payload in Byte, as well as the Transaction ID (32 bit), to identify related messages as request and reply.

Figure 3.20 shows a first practical example with a Hello message ((1) in Figure 3.18) with which the switch reports to the controller to establish an OpenFlow connection and informs it in the payload field element of the supported OpenFlow version 1.4 using the bitmap value 00000020. On request, the switch informs the controller with Features Reply ((4) in Figure 3.18, Type OFPT_FEATURES_REPLY), according to Figure 3.21, about the capabilities it supports. The datapath_id identifies the switch (comparable to a bridge MAC address), n_buffers the number of input queues for Packet-in packets (here 256), n_tables the number of supported flow tables (here 254). Capabilities describe supported functions such as collecting statistics for flows, tables, ports, groups, etc.

**Fig. 3.19:** Structure of the header of an OpenFlow message

```
OpenFlow 1.4
Version: 1.4 (0x05)
Type: OFPT_HELLO (0)
Length: 16
Transaction ID: 4
▼ Element
  Type: OFPHET_VERSIONBITMAP (1)
  Length: 8
  Bitmap: 00000020
```

Fig. 3.20: Hello OpenFlow message (1) captured with protocol analysis software

```
OpenFlow 1.4
Version: 1.4 (0x05)
Type: OFPT_FEATURES_REPLY (6)
Length: 32
Transaction ID: 2843106426
datapath_id: 0x0000000000000001
n_buffers: 256
n_tables: 254
auxiliary_id: 0
Pad: 0
▼ capabilities: 0x0000004f
  .... .... .... .... .... ...1 = OFPC_FLOW_STATS: True
  .... .... .... .... .... ...1.. = OFPC_TABLE_STATS: True
  .... .... .... .... .... .1... = OFPC_PORT_STATS: True
  .... .... .... .... .... 1... = OFPC_GROUP_STATS: True
  .... .... .... .... .... ..0. .... = OFPC_IP_REASM: False
  .... .... .... .... .... .1... .... = OFPC_QUEUE_STATS: True
  .... .... .... .... .... ..0 .... .... = OFPC_PORT_BLOCKED: False
Reserved: 0x00000000
```

Fig. 3.21: Features Reply OpenFlow message (4) captured with protocol analysis software

Figure 3.22 represents a Packet-in OpenFlow message ((6) in Figure 3.18, Type OFPT_PACKET_IN) from network practice. The reason for this message sent from the switch to the controller is an ICMP (Internet Control Message Protocol) packet received via port 1, for which there is no flow table entry yet (reason: OFPR_TABLE_MISS). After evaluation in the controller, the controller makes the necessary additional settings in the switch with a Modify-State message according to Figure 3.23 ((7) in Figure 3.18, type OFPT_FLOW_MOD, Command ADD). In the example, the input port 2 (IN_PORT) and the Ethernet target address 00:00:00:00:00:01 (ETH_DST) must be set for the match fields so that packets belonging to this particular flow are sent out via port 1 (action OUTPUT).

```

OpenFlow 1.4
Version: 1.4 (0x05)
Type: OFPT_PACKET_IN (10)
Length: 140
Transaction ID: 0
Buffer ID: 258
Total length: 98
Reason: OFPR_TABLE_MISS (0)
Table ID: 0
Cookie: 0x0000000000000000
▼ Match
  Type: OFPMT_OXM (1)
  Length: 12
  ▼ OXM field
    Class: OFPXMC_OPENFLOW_BASIC (0x8000)
    0000 000. = Field: OFPXMT_OFB_IN_PORT (0)
    .... ...0 = Has mask: False
    Length: 4
    Value: 1
    Pad: 00000000
  Pad: 0000
▼ Data
  > Ethernet II, Src: 00:00:00_00:00:01 (00:00:00:00:00:01), Dst: 00:00:00_00:00:02 (00:00:00:00:00:02)
  > Internet Protocol Version 4, Src: 10.0.0.1, Dst: 10.0.0.2
  > Internet Control Message Protocol

```

Fig. 3.22: Packet-in OpenFlow message (6) captured with protocol analysis software

```

OpenFlow 1.4
Version: 1.4 (0x05)
Type: OFPT_FLOW_MOD (14)
Length: 96
Transaction ID: 2843106431
Cookie: 0x0000000000000000
Cookie mask: 0x0000000000000000
Table ID: 0
Command: OFPFC_ADD (0)
Idle timeout: 0
Hard timeout: 0
Priority: 1
Buffer ID: OFP_NO_BUFFER (4294967295)
Out port: 0
Out group: 0
Flags: 0x0000
Importance: 0
Match
  Type: OFPMT_OXM (1)
  Length: 22
  ▼ OXM field
    Class: OFPXMC_OPENFLOW_BASIC (0x8000)
    0000 000. = Field: OFPXMT_OFB_IN_PORT (0)
    .... ...0 = Has mask: False
    Length: 4
    Value: 2
  ▼ OXM field
    Class: OFPXMC_OPENFLOW_BASIC (0x8000)
    0000 011. = Field: OFPXMT_OFB_ETH_DST (3)
    .... ...0 = Has mask: False
    Length: 6
    Value: 00:00:00_00:00:01 (00:00:00:00:00:01)
    Pad: 0000
Instruction
  Type: OFPIT_APPLY_ACTIONS (4)
  Length: 24
  Pad: 00000000
Action
  Type: OFPAT_OUTPUT (0)
  Length: 16
  Port: 1
  Max length: 65509
  Pad: 000000000000

```

Fig. 3.23: Modify-State OpenFlow message (7) captured with protocol analysis software

The SDN concept was initially developed by the Open Networking Foundation (ONF) [148] and standardized in version 1.0.0, including the Southbound API control protocol OpenFlow [147]. Meanwhile, there are more OpenFlow versions with extensions and improvements. The latest version is 1.5.1 [149]. But there are also alternatives to OpenFlow. To be mentioned here are, for example, NETCONF (Network Configuration Protocol), OVSDB (Open vSwitch Database Management Protocol), BGP-FS (Border Gateway Protocol-Flow Spec), PCEP (Path Computation Element Communication Protocol), OpenConfig, XMPP (Extensible Messaging and Presence Protocol), or I2RS (Interface to the Routing System) [89].

Interestingly, although there are several specifications for the southbound API, the D-CPI in Figure 3.17, there is no standard for the northbound API, the A-CPI

(Applications-Controller Plane Interface). In practice, however, a RESTful API (Representational State Transfer) is usually used here [89]. As shown in Figure 3.17, SDN controllers can also exchange information with other network domains for routing optimization purposes, usually via routing protocols such as BGP (Border Gateway Protocol). If the communication takes place with another SDN network or controller, this is called a Westbound API. If it is a legacy network, for example, with MPLS routers, the interface is called Eastbound API [173].

Considering the comments made above on SDN, it is now obvious why SDN is suitable for supporting service chains. Due to the adaptability of the flow tables in the SDN switches, a complete VNF-FG can be configured as required via an SDN controller and also adapted at any time since, for example, the modification of the IP destination address according to the next VNF to be passed through in the FG can be specified as an action. The SDN controller, in turn, receives its preferences from the corresponding NFV SDN application. The parameters for this can be provided to the SDN/network controller by a VIM in the NFV MANO, as shown in Figure 3.24 [132] (see Section 3.1, Figure 3.4). Concerning the transport network and the desired service chains, SDN represents a practical solution for flexible flow handling in an NFV environment. Different approaches to the interaction of NFV and SDN, in addition to the solution shown in Figure 3.24, can be found in [97].

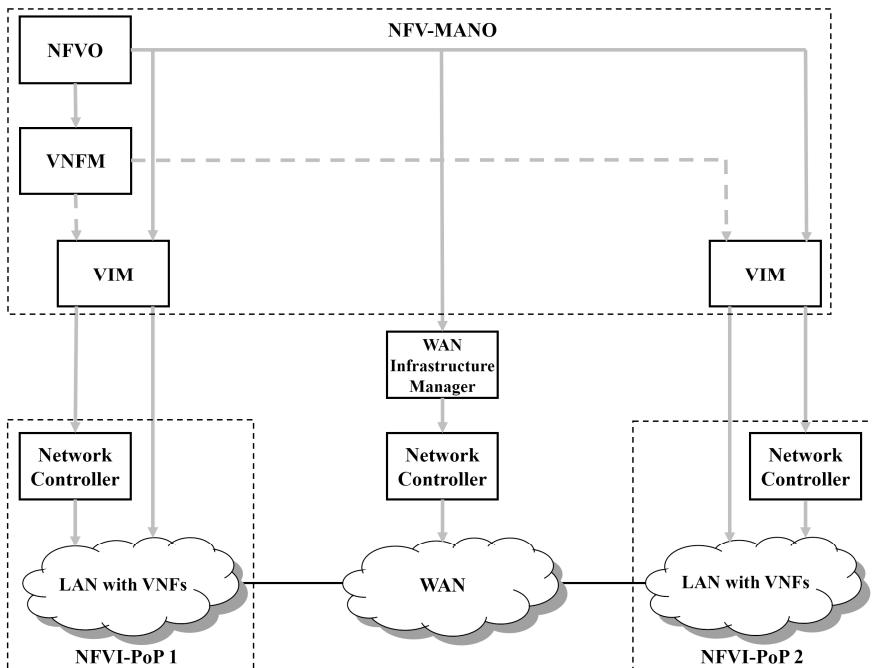


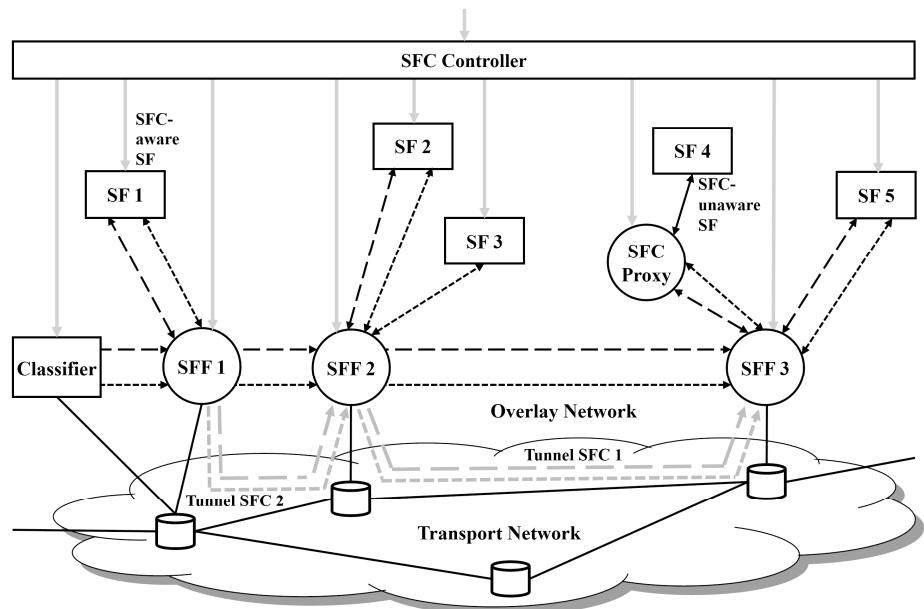
Fig. 3.24: NFV MANO and SDN controller [132]

Besides the support of service chains, which are very important for modern networks, there are numerous other use cases where SDN can be used to benefit:

- Cloud orchestration: Management of the servers in the cloud and associated networks can be integrated with SDN. Depending on the migration of the VMs (number and/or location), the transport network can be reconfigured automatically. Conversely, if the links are overloaded, VMs can be moved to a more suitable location.
- Load Balancing: Each SDN switch can be controlled from the SDN controller as a load balancer so that, depending on the load, e.g., service requests from clients are forwarded to different servers providing the same service. Additional load balancer network elements are not required.
- Routing modifications: Due to the central view of the network, changes regarding path selection, traffic optimization, redundant paths, different protocol versions (e.g., IPv4, IPv6), and routing protocols are much more comfortable than with monolithic routers.
- Traffic monitoring and measurements: In an SDN network, information on the status of the network is collected centrally by definition, which is then, of course, also available for measurement and evaluation purposes. Also, an SDN infrastructure provides monitoring access to any packet flow of interest without additional effort (no network taps required), e.g., to determine delay times.
- Network management: In legacy networks, the policies (e.g., access control lists) must be configured for each network node at high effort. The central control in an SDN network simplifies this. An automated and optimized adaptive setting is possible.
- Application-specific network optimization: Using one of the SDN controller's Northbound APIs, an SDN application can inform the transport network about its properties and status. Corresponding forwarding decisions or resources can be requested. Conversely, the SDN controller can communicate its network view to the application and induce it to change its behavior, e.g., in case of a resource bottleneck.
- Test networks for research (e.g., new routing protocol), prototypical implementations in the development and rollout of new SW and protocol versions
- Parallel operation of several virtual transport networks, if required, with separate SDN controllers, for different areas of application (e.g., for 1. telecommunications, 2. smart grid, 3. testing of new releases) based on the same hardware. In such a case, the term Network Slices is used [173].

Another, still relatively new solution to the service chaining problem has been proposed by the IETF and standardized in several RFCs [15; 17; 18] under the headline Service Function Chaining (SFC). As Figure 3.25 shows, the service chains consist of the (virtual) network functions SF 1 (Service Function) to n. A required network service is formed by concatenation, i.e., a sequential combination of SFs. Figure 3.25

shows the two exemplary Service Function Chains (SFC) 1 and 2 with the SFs 1 – 2 – 4 – 5 and 1 – 2 – 3 – 4 – 5. The passing through such a service sequence in the network is called Service Function Path (SFP). The delivery of a packet or frame to one SF and the forwarding to the next SF is done by the logical function Service Function Forwarder (SFF). If an SF can process the concatenation information, this is called an SFC-aware SF. If it is unable to do so, it is an SFC-unaware SF. In this case, a gateway function in the form of a so-called SFC proxy must be switched between SFF and SFC-unaware SF. The required SFCs are defined by configuring the corresponding SFs, SFC proxies, and the SFFs accordingly via an SFC controller. The classifier is also configured for the specific classification criteria (policies) of each SFC. Each incoming IP packet or Ethernet frame is assigned to the appropriate SFP according to the classification criteria and subsequently routed from SF to SF through the SFP by the SFFs. Together, these network elements for the service chains form an overlay network with tunnels for each SFC [15; 17; 73].



Network Service 1 => SFC 1 => SFP 1: SF1 → SF2 → SF4 → SF5 = ——

Network Service 2 => SFC 2 => SFP 2: SF1 → SF2 → SF3 → SF4 → SF5 = -----

SF = Service Function

SFF = Service Function Forwarder

SFC = Service Function Chain

SFP = Service Function Path

Fig. 3.25: SFC architecture

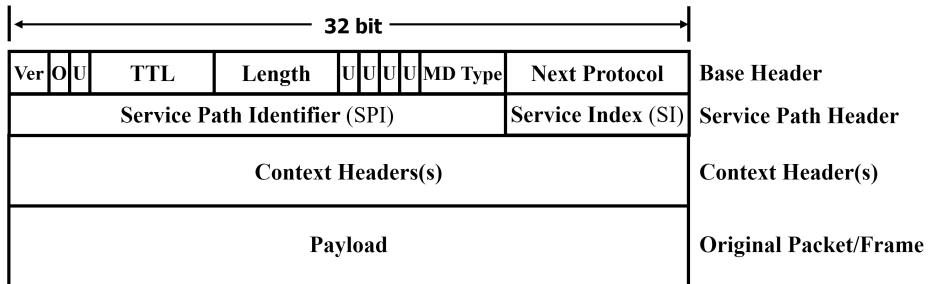
A specific SFC must be uniquely defined. Therefore a Service Path Identifier (SPI) is introduced. Moreover, the correct sequence of SFs within an SFP must be guaranteed. A Service Index (SI) ensures this, counting down from 255. I.e., in the above example with the SFs 1 – 2 – 3 – 4 – 5, we start with SI = 255 and then count down to SI = 251. Furthermore, sometimes meta and control data have to be passed on from one SF to another SF. A so-called Network Service Header (NSH), according to RFC 8300 [18], is introduced for this additional information required in the context of a service chain. Figure 3.26 shows its structure. An NSH is organized in lines of 32 bit each. The first line contains the so-called base header. It consists of a 2-bit version field with the current value 0, an O-bit which is set to 1 for OAM data (Operation, Administration and Maintenance), otherwise, to 0, a 6-bit TTL value (Time To Live) decremented from the default value 63 from each SFF to prevent endless loops – if TTL = 0 the packet or frame is discarded – and a 6-bit length field which specifies the total SH length in 32-bit lines. The 4-bit field MD Type specifies whether the context header has a fixed or variable length. Finally, the 8-bit field Next Protocol specifies which type of protocol message is transported in the payload field: IPv4 (01 hex), IPv6 (02 hex), Ethernet (03 hex), NSH (04 hex), or MPLS (05 hex). Furthermore, the base header contains U-bits reserved for future purposes (Unassigned).

Essential for a service chain is the service path header. It consists of a unique 24 bit SPI and the 8 bit SI, which is decremented by 1 per SF when passing through a chain.

From one SF to the next SF in an SFP, specific data, which depend on the way the SF is executed, may have to be passed on as a result. This metadata, “data about data”, is transported in one or more context header fields. This is used to pass information in the SFC from one SF(i) (or classifier) to the next SF(i+1), which is only available at SF(i) or can only be obtained at SF(i+1) with much effort, but is needed at SF(i+1). An example of this is the IP network element S-GW in the EPC of an LTE network (see Section 2.5, Figure 2.29). An S-GW as an IP edge router receives encapsulated IP packets from an eNodeB as the tunnel endpoint and subsequently determines the corresponding Subscriber ID and the policies (processing criteria) for this IP flow by querying the PCRF using the Diameter protocol. This information is essential for the further handling of IP packets in the service chain, but further inside the network, it is difficult or impossible to determine [156].

Figure 3.27 shows a practical example with an NSH encapsulated Ethernet II frame with the SPI 39030 and an SI of 253. The latter means that this overlay frame was captured at the third SF of the SFC.

The NSH processing in Figure 3.25 is as follows: The classifier or an SFC proxy inserts an NSH. An SFF forwards an NSH. Here the mapping of SPI and SI to a real next hop (IP or MAC address) and transport protocol for tunneling (e.g., VXLAN, GRE, MPLS) takes place. The last SFF or SFC proxy in a chain removes the NSH. An SF or an SFC proxy decrements the SI and updates the context header if necessary [18; 15; 17].

**Fig. 3.26:** Structure of the Network Service Header (NSH)

```

Frame 111: 609 bytes on wire (4872 bits), 609 bytes captured (4872 bits) on interface 0
Linux cooked capture
Network Service Header
  00.. .... .... .... = Version: 0 (0x0)
  ..0. .... .... .... = O Bit: 0
  ...0 .... .... .... = C Bit: 0
  .... 1111 11.. .... = Reserved Bits: 0x3f
  .... .... ..00 0110 = Length: 6 (0x06)
  MD Type: 1 (0x01)
  Next Protocol: Ethernet (3)
  SPI: 39030 (0x0009876)
  SI: 253 (0xfd)
  Context Header: 00000000
  Context Header: 00000000
  Context Header: 00000000
  Context Header: 00000000
Ethernet II, Src: Xerox_00:00:10 (00:55:00:00:00:10), Dst: 00:00:00_aa:00:02 (00:00:00:aa:00:02)
Internet Protocol Version 4, Src: 10.0.6.10, Dst: 10.0.1.20
Transmission Control Protocol, Src Port: 80, Dst Port: 39960, Seq: 1, Ack: 412, Len: 503
Hypertext Transfer Protocol
Line-based text data: text/html (9 lines)

```

Fig. 3.27: Encapsulated Ethernet frame with NSH header captured with protocol analysis software

In the transport network shown in Figure 3.25, the original packets or frames are each extended by one NSH and tunneled from SF to SF encapsulated in an Ethernet frame or IP packet.

It should also be mentioned at this point that SDN can not only be used as an alternative to the SFC discussed but that SDN can also be utilized as a technical basis for the implementation of SFC. Furthermore, the interaction between NFV MANO and an SFC controller in Figure 3.25 can be carried out as for SDN in Figure 3.24: The SFC controller can be managed via a VIM from the NFV MANO, i.e., it is informed of the SFCs currently required.

3.3 Future Networks Concept

The NFV and SDN techniques discussed in Sections 3.1 and 3.2 are an essential part of the standardization activities carried out at ITU-T on the so-called Future Networks. Recommendation Y.3001 [179] defines a Future Network (FN) in quite general terms “A network able to provide services, capabilities, and facilities difficult to provide using existing network technologies”. In more concrete terms, [179] describes FN in terms of four objectives and twelve corresponding design goals, which are explained briefly in the following and are shown in Figure 3.28.

The four objectives focus on network aspects that have been little or not considered in previous networks, including NGNs (see Section 1.3):

- Service awareness: FN should provide current and future services tailored to the needs of applications and users. This means, among other things, that a massive number of different services can be offered at moderate deployment and operating costs.
- Data awareness: FN should provide a network architecture that can handle vast amounts of data in a distributed environment. Users should be able to retrieve desired data securely, easily, quickly, and reliably regardless of their location, whereby the term data refers to all information that can be accessed via a network in addition to audio and video.
- Environmental awareness: FN should be environmentally friendly, i.e., architecture, implementation, and operation should ensure a minimum impact of the network on the environment, in particular minimizing material and energy consumption and greenhouse gas emissions. Furthermore, an FN should support other industries in their environmental sustainability.
- Social and economic awareness: FN should consider social and economic issues so that the entry barriers for different actors are reduced. These include reducing lifecycle costs, providing access to the network and services regardless of location, and enabling competition and financial revenues.

In the past, the latter two objectives, in particular, were not *a priori* important for networks, especially as developments were technology-driven. The specific focus on data is due to the development of M2M communications and the IoT. Services have already played a significant role in the NGN concept [173].

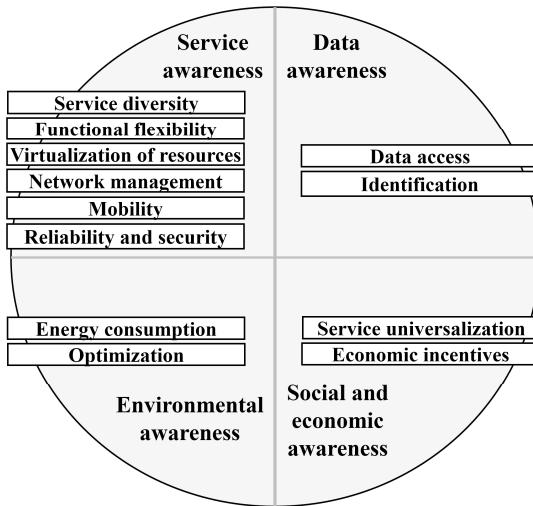


Fig. 3.28: Four objectives and twelve design goals for Future Networks [179]

As already mentioned above and shown in Figure 3.28, twelve design goals were derived in [179] – based on the four objectives explained above – which characterize Future Networks in more detail:

- Service diversity: support of a wide range of services with different traffic characteristics (bit rate, latency) and behavior (security, reliability, mobility) and a large variety of end systems (e.g., from high-resolution video conferencing systems to simple sensors)
- Function flexibility: flexible support (e.g., video transcoding, sensor data aggregation, new protocols), including the agile provision of new services in response to unforeseen user requests
- Virtualization of resources: virtualize the physical resources and introduce an abstraction layer and provide different virtual networks that work independently of each other
- Network management: efficient, automated operation, monitoring, and provisioning of services and network facilities despite a massive amount of management data
- Mobility: mobility support of a massive number of end systems, which may move at high speed across heterogeneous networks (e.g., different access networks, radio cell sizes)
- Reliability and Security: network design and operation in terms of high availability and adaptability (e.g., for emergency and disaster scenarios, management of road, rail, and air traffic, smart grids, medical care) as well as security and privacy for users

- Data access: efficient handling of large amounts of data (e.g., in social networks or sensor networks) and provision of mechanisms for fast access to data regardless of the location where the data is provided (efficient storage and quick search mechanisms)
- Identification: provision of new identification methods (beyond the IP addresses commonly used today) for effective and scalable mobility support as well as data access
- Energy consumption: improving energy efficiency and saving energy in all areas, meeting user requirements with minimal network traffic
- Optimization: adaptation of the network performance and corresponding optimization of the network equipment to the real service and user requirements, not to the maximum possible requirements as in current networks
- Service universalization: enable service deployment in urban or rural areas, developed or developing countries by reducing life cycle costs and using global standards
- Economic incentives: provide a sustainable and competitive environment (e.g., without the lack of QoS support as in today's IP networks) for different stakeholders such as users, providers, government institutions, and rights holders.

In addition to the characteristics mentioned above, [179] also points out possible technologies to achieve these design goals, especially virtualization with NFV [180] (see Section 3.1). Figure 3.29 shows the importance of virtualization for Future Networks and explains the relationships. The physical networks or resources (networks, computers, and storage resources) are partitioned and abstracted as virtual resources (virtual machines, virtual network functions). The latter form the basis for creating virtual networks, so-called LINPs (Logically Isolated Network Partitions), which implement service-specific networks. This means that the LINPs for different services can be considered in isolation. A physical resource can be shared among many virtual resources, while a LINP is composed of many virtual resources. Further details of such a Future Network can be found in [181]. This FN standard describes SDN (see Section 3.2) as the key technology for Future Networks [173].

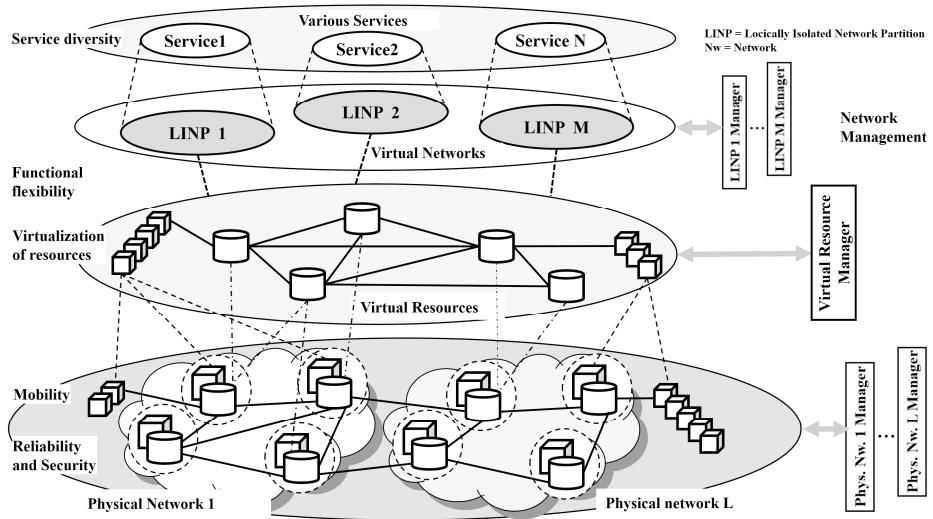


Fig. 3.29: Virtualization in Future Networks [180]

In summary, we can state that the Future Networks approach standardized by ITU-T and outlined here is an essential basis for 5G. This becomes obvious when comparing the FN network architecture in Figure 3.29 with the targeted 5G network architecture in Section 6.3.

4 5G Use Cases and Requirements

The procedure on the way to 5G differed and still differs significantly from that of previous generations of mobile networks, including 3G and 4G. While in the past, the focus was on communication between and services for people, it is now on providing a networked world for everyone and everything, i.e., not only for people but also for (smart) things and systems. The approach is no longer primarily technology-driven, like up to and including 4G, but use case-driven. Based on a large number of possible use cases, the requirements were derived, and the technology required for implementation was specified [88]. This process took place in several projects and organizations. In chronological order, we should mention the following: the EU research project METIS (Mobile and wireless communications Enablers for the Twenty-twenty Information Society) [139], the NGMN Alliance (Next Generation Mobile Networks) [143], ITU-R (ITU-Radiocommunication Sector) [115], 5GPPP (5G Infrastructure Public Private Partnership) [66] and last but not least 3GPP.

The primary research project METIS was part of the European 5G funding. In general, EU 5G research funding has taken place and continues to take place under the umbrella of the 5G PPP, a joint European initiative. In this context, it is also important to mention related activities in other parts of the world: in the USA, the 5G Americas organization [272] and the TIA (Telecommunications Industry Association) [270], in China, the IMT-2020 Promotion Group [273], in South Korea, the 5G Forum [274], and in Japan, the 5GMF (Fifth Generation Mobile Communications Promotion Forum) [275]. These efforts have been and are being supplemented by many research projects on 5G at universities, in companies, and in joint projects worldwide.

4.1 5G Use Cases and Usage Scenarios

The METIS project published scenarios, use cases (here called test cases), and the first requirements for a 5G network in April 2013 [98]. Table 4.1 provides an overview of them. In summary, the following main requirements were formulated [151; 173]:

- 1 to 10 Gbit/s data rate, e.g., for a virtual reality office
- 9 GByte/hour data volume in the busy period, e.g., in a stadium, and 500 GByte/month and subscriber in a dense urban information society
- Less than 5 ms end-to-end latency, e.g., for traffic efficiency and safety
- 10 years battery lifetime for applications with massive deployment of sensors and actuators
- 300000 devices per access point
- 99,999% reliability, e.g., for teleprotection in smart grid and traffic efficiency and safety

- Energy consumption as in 4G
- Similar costs as for 4G.

This list already shows that not all of these ambitious requirements have to be met for every application in Table 4.1 and that 5G network technology must be very flexible, not only in terms of costs and energy consumption.

Tab. 4.1: Scenarios, essential requirements, and use cases in the EU research project METIS [98]

Use/test case	Scenario				
	Amazingly fast in a crowd	Great service in a crowd	Ubiquitous things com- municating	Best experi- ence follows you	Super real-time and reliable connections
Essential requirements					
Very high data rate	Very dense crowds of users	Very low ener- gy, cost, and a massive num- ber of devices	Mobility	Very low latency	
Virtual reality office	X				
Dense urban information society	X	X	X	X	
Shopping mall		X	X		
Stadium		X			
Teleprotection in smart grid			X		X
Traffic jam	X			X	
Blind spots in rural and urban areas				X	
Real-time remote compu- ting for mobile terminals				X	X
Open-air festival	X		X		
Emergency communica- tions			X	X	X

Massive deployment of sensors and actuators	X	X	X
Traffic efficiency and safety		X	X

The NGMN Alliance is an association of major mobile network operators, manufacturers and research institutes to promote and influence the standardization of future mobile networks. They published in February 2015 a highly regarded white paper [96] with 14 application categories and 24 use cases with correspondingly derived requirements for a 5G network. Table 4.2 provides an overview.

Tab. 4.2: Use cases and categories from NGMN [96]

Category	Use case	Requirements			
		User experienced data rate	End-to-end latency	Mobility	Connection density
Broadband access in dense areas	– Pervasive video	DL (Down): 300 Mbit/s	10 ms	0-100 km/h	200-2500/km ²
	– Cloud services	UL (Up Link):			
	– Dense urban society	50 Mbit/s			
Indoor ultra-high broadband access	– Smart office	DL: 1 Gbit/s	10 ms	Pedestrian	75000/km ²
		UL:			
		500 Mbit/s			
Broadband access in a crowd	– HD video/photo sharing in a stadium or open-air gathering	DL: 25 Mbit/s	10 ms	Pedestrian	150000/km ² , 30000/stadium
		UL:			
		50 Mbit/s			
50+ Mbit/s everywhere	– 50 Mbit/s everywhere	DL: 50 Mbit/s	10 ms	0-120 km/h	400/km ² in suburban, 100/km ² in rural
		UL:			
		25 Mbit/s			
Low-cost broadband	– Ultra-low-cost networks	DL: 10 Mbit/s	50 ms	0-50 km/h	16/km ²
		UL:			
		10 Mbit/s			
Mobile broadband in vehicles (cars, trains)	– High-speed trains	DL: 50 Mbit/s	10 ms	up to 500 km/h	2000/km ² , 500 active users per train, 4 trains
	– Moving communication hotspots	UL:			
	– Remote computing	25 Mbit/s			

Airplanes connectivity	– 3D connectivity for aircrafts	DL: 15 Mbit/s UL: 7,5 Mbit/s	10 ms	up to 1000 km/h	60 airplanes/ 18000 km ²
Massive low-cost, long-range, low-power MTC	– Smart wearables and clothing – Sensor networks	≤ 100 kbit/s	s - h	0-500 km/h	200000/km ²
Broadband MTC (Machine Type Communications)	– Mobile video surveillance	DL: 50 Mbit/s UL: 25 Mbit/s	10 ms	0-120 km/h	200-2500/km ²
Ultra-low latency	– Tactile internet	DL: 50 Mbit/s UL: 25 Mbit/s	< 1 ms	Pedestrian	not critical
Resilience and traffic surge	– Natural disaster	≤ 1 Mbit/s	not critical	0-120 km/h	10000/km ²
Ultra-high reliability and ultra-low latency	– Automatic traffic control and driving – Collaborative robots – Remote object manipulation	≤ 10 Mbit/s	1 ms	0-500 km/h	not critical
Ultra-high availability and reliability	– eHealth for life-critical applications – Public safety – 3D connectivity, e.g., for drones	10 Mbit/s	10 ms	0-500 km/h	not critical
Broadcast like services	– News and information – Local, regional, or national broadcast-like services	DL: ≤ 200 Mbit/s UL: ≤ 500 kbit/s	< 100 ms	0-500 km/h	not critical

In summary, the NGMN 5G view, according to [96], can be described as follows:

- User Experience: 1 Gbit/s, e.g., indoor, at least 50 Mbit/s everywhere; 1 ms end-to-end latency for tactile communication; very high mobility requirements, e.g., for high-speed trains, but also stationary operation, e.g., of smart meters
- System-Performance: several 10 Mbit/s per user for several 10000 users, e.g., in a stadium; 1 Gbit/s per user for up to 10 users, e.g., in an office; several 100000 simultaneous connections per km² for massive scaled sensors; improved spectral efficiency compared to 4G, higher coverage in rural areas and more efficient signaling due to energy consumption

- Device Requirements: high degree of programmability and configurability; operation in different frequency ranges and modes; traffic aggregation with simultaneous use of several radio technologies; operation of low-cost MTC devices; increased battery lifetime of at least 3 days for a smartphone and up to 15 years for an MTC device, e.g., a sensor
- Enhanced Services: seamless connection despite different access points and RAT networks (Radio Access Technology); position accuracies of less than 1 m in 80% of cases, and inside rooms; high network security and guaranteed privacy despite heterogeneous access networks; high availability, for specific use cases up to 99.999%
- New Business Models: for connectivity provider, service provider, 3rd party service provider, and XaaS asset provider (X as a Service) as well as a shared network for several network operators and verticals (network sharing)
- Network Deployment, Operation, and Management: 1000 times more traffic than today with half the energy consumption; configuration options with the aim of low energy consumption or high performance; easy integration of new services and new RATs; high flexibility and scalability; fixed-mobile convergence with consistent user experience; low operational costs [96; 173].

These results also show that due to the different requirements in different use case scenarios, a 5G network has to provide a wide range of services at various locations with widely varying bit rates and numbers of connected terminals. This requires enormous flexibility, scalability, and elasticity [173].

In September 2015, the ITU-R published its groundbreaking IMT (International Mobile Telecommunications) vision for 2020 and beyond with the Recommendation M.2083: IMT-2020 [128]. Usage scenarios were developed and summarized in three main areas:

- Enhanced Mobile Broadband (eMBB)
- Ultra-Reliable and Low Latency Communications (URLLC)
- Massive Machine Type Communications (mMTC).

Enhanced Mobile Broadband addresses the human-centric use cases for access to multimedia content, services, and data. This includes personal communications, on the one hand, at mobile hotspots with high user density, high bit rates but low mobility, and on the other hand, in a wider geographical area with lower bit rate requirements but uninterrupted connectivity even at high mobility.

Ultra-Reliable and Low Latency Communications have stringent requirements for capabilities such as throughput, latency, and availability. Examples include wireless control of industrial manufacturing or production processes, remote medical surgery, distribution automation in a smart grid, transportation safety, etc.

Massive Machine Type Communications is characterized by a very large number of connected devices, typically transmitting a relatively low volume of non-delay

sensitive data. Devices are required to be low cost and to provide a very long battery life.

Figure 4.1 shows the three 5G usage scenarios in a triangular arrangement highlighting the different requirements. Here we will also find some use cases not yet mentioned, such as Smart City, which due to their arrangement in a triangle, also show overlaps and transitions in the three described scenarios. This results in extreme demands on 5G technology, which never have to and can be met in total but only in parts according to the use case. These require a modular design for 5G networks [128].

The requirements resulting from the consideration of these three usage scenarios and their possible combinations and overlaps are subject to Section 4.3.

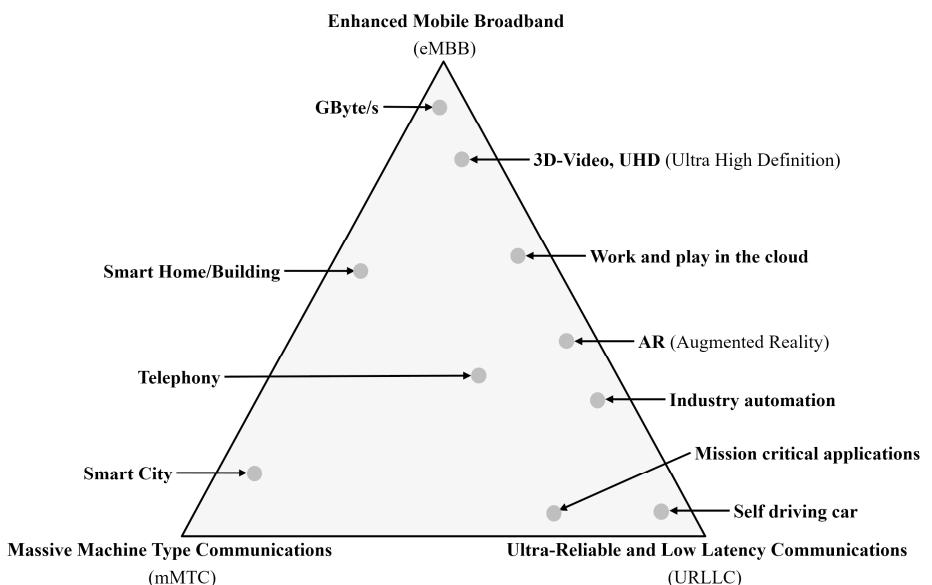


Fig. 4.1: ITU-R usage scenarios of IMT-2020 [128]

Simultaneously to the development of the ITU vision, numerous large research projects on 5G were in progress worldwide. In Europe, these were brought together under the 5GPPP umbrella and include the EU projects initiated by the European Commission and the ICT industry (Information and Communication Technology) like METIS II, FANTASTIC-5G, mmMAGIC, SPEED-5G, 5G-NORMA, Flex5GWARE, and VirtuWind, etc. In this context, 5GPPP maintained a document entitled “5G PPP use cases and performance evaluation models” [134]. In version 1.0 of April 2016, the use cases from the EU research projects mentioned in the context of 5GPPP were collected and structured in six scenarios, here called use case families:

- Dense urban
- Broadband (50+ Mbit/s) everywhere
- Connected vehicles
- Future smart offices
- Low bandwidth IoT
- Tactile internet/automation.

These show significant similarities with the METIS, NGMN, and ITU-R results. In parallel, [134] also describes an industry-driven approach with the sectors (vertical industries)

- Automotive,
- eHealth,
- Energy,
- Media and entertainment,
- Factories of the future

including the assigned business cases.

Both approaches are mapped to each other so that the requirements from the use cases are also available for the industries with their business cases.

In September 2016, 3GPP adopted as part of 3GPP Release 14 a study in the form of TR 22.891 [27]. This document summarizes 74 use cases in five categories based on previous results, experience, and standardization work:

- Enhanced Mobile Broadband (eMBB): Examples of use cases are mobile broadband communication, UHD television (Ultra High Definition), hologram, augmented reality, virtual reality, high mobility in trains or airplanes, virtual presence.
- Critical Communications (CriC): e.g., interactive games, sports broadcasts, industrial control, drones, robots. ITU-R lists this category under “Ultra-Reliable and Low Latency Communications (URLLC)” [128].
- Massive Machine Type Communications (mMTC), Massive Internet of Things (MIoT): use cases in metro or stadium, eHealth, smart city (eCity), smart farming (eFarm), wearables, inventory control
- Network Operation: e.g., network slicing, routing, migration, and interworking, energy saving
- Enhancement of Vehicle-to-Everything (eV2X): e.g., autonomous driving.

Figure 4.2 shows an overview of the above categories and use cases [27]. In comparison with the results of ITU-R and the three resulting usage scenarios, there are two additional use case categories, one for the network itself and one for the crucial V2X (Vehicle-to-Everything) application cases.

TR 22.891 already contains references to the requirements for each use case. For the four categories relevant to end users they are summarised in Table 4.3. The re-

quirements are then discussed in more detail in Section 4.3, based on further standards.

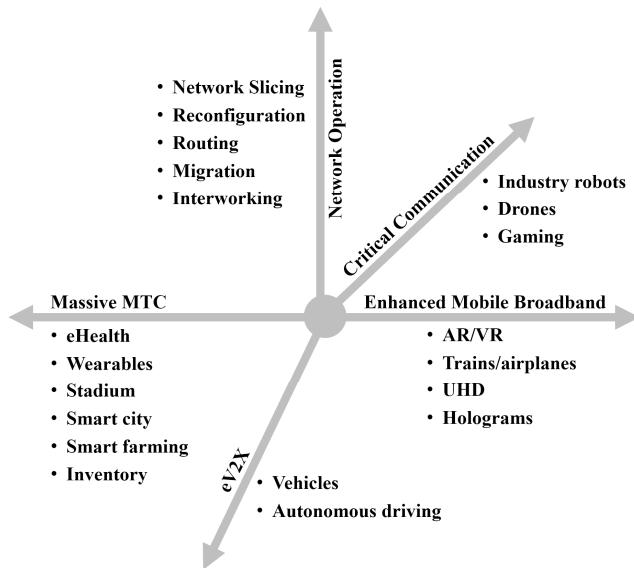


Fig. 4.2: 3GPP usage categories and examples of use cases for 5G [27]

Tab. 4.3: 3GPP use case categories and resulting general requirements [27]

Category	General requirements
eMBB (Enhanced Mobile Broadband)	<ul style="list-style-type: none"> – Very high data rate, up to 10 Gbit/s per user – Low latency – High traffic density, Tbit/s/km² – High density for UE, up to 2500 UEs/km² – Mobility from 0-500 km/h – No special requirements for availability and position accuracy
CriC (Critical Communications) or URLLC (Ultra-Reliable and Low Latency Communications)	<ul style="list-style-type: none"> – No special requirements for data rate and mobility – Very low latency, < 1 ms end-to-end – Ultra-high reliability, high availability – High density, < 1000 UEs (e.g., sensors)/km² – Precise position, ≤ 10 cm
MIoT (Massive Internet of Things) or mMTC (Massive Machine Type Communications)	<ul style="list-style-type: none"> – No special requirements for data rate, latency, reliability, and mobility – Efficient communication to support devices with limited resources and low power battery supply – Very high density, up to 1 Mio. UEs (e.g., sensors)/km² – High positioning accuracy, ≤ 50 cm

Category	General requirements
eV2X (Enhancement of Vehicle-to-Everything)	<ul style="list-style-type: none"> – Medium data rate, 10 Mbit/s per device – Very low latency, ≤ 1 ms end-to-end – Ultra-high reliability, nearly 100 % – Medium traffic density – Medium connection density, > 10000 vehicles on a road with several lanes – High mobility, up to 500 km/h – Precise position, ≤ 10 cm

4.2 Application Areas for 5G

Looking at the use cases for 5G networks mentioned in Section 4.1, it is immediately apparent that 5G networks are used in many areas of life. [88] provides an incomplete but illustrative overview of applications and industries significantly affected by 5G. The list shows possible use cases for each area of application for 5G:

- Manufacturing: e.g., for remote or motion control and monitoring of devices like robots, machine-to-machine communication, Augmented Reality (AR), and Virtual Reality (VR) in design (e.g., for designing machines, houses, etc.)
- Automotive: for example, for platooning, e.g., with trucks, infotainment, autonomous vehicles, high-resolution map updates, remote maintenance, and SW updates
- Entertainment: e.g., for mobile UHD video streaming (Ultra High Definition), stadium experience, VR, cooperative media production (e.g., production of songs, movies from various locations)
- Energy: for example, for grid control and monitoring, connecting wind farms, smart electric vehicle charging
- Public transport: Example use cases are infotainment, train or bus operations, and platooning for buses.
- Agriculture: e.g., for connecting sensors and farming machines, drone control
- Public Safety: e.g., threat detection, facial recognition, drone control
- Healthcare: e.g., for bioelectronic medicine, personal health systems, telemedicine, connected ambulance including AR/VR applications
- Fixed Wireless Access (FWA): replacing fixed access technologies like fiber at the last mile by 5G wireless access
- Megacities: applications around mission control for public safety, video surveillance, connected mobility across all means of transport, including public parking and traffic steering, and environment or pollution monitoring.

Special efforts to shape the development towards 5G have been made in the areas of manufacturing – often described by the term Industry 4.0, or Industrial Internet –

and automotive. In these application areas, there are many highly interesting use cases with requirements that cannot be met by previous mobile networks. Therefore, we will examine these two areas of application in more detail.

5G and Industry 4.0 were the main topics of the 5G-ACIA (5G Alliance for Connected Industries and Automation). In this alliance, numerous well-known companies have joined together as a working group within the ZVEI (Zentralverband Elektrotechnik- und Elektronikindustrie e.V.) to ensure the best possible applicability of 5G technology for industry, in particular the manufacturing and process industry [70].

Figure 4.3 illustrates exemplary applications for a 5G network in the factory of the future. Various examples are presented of how the advantages of 5G can be used in a factory. Applications range from logistics for supply management, networking of AGVs (Automated Guided Vehicle, autonomous transport robot) and sensors, control of an assembly line, motion control and cooperation of production robots, localization of devices and articles, to inventory management and logistics for delivery. Additionally, this wireless networking, which meets the communication requirements of Industry 4.0, is not limited to the LAN, the individual factory site but is available across WANs in the production process end-to-end.

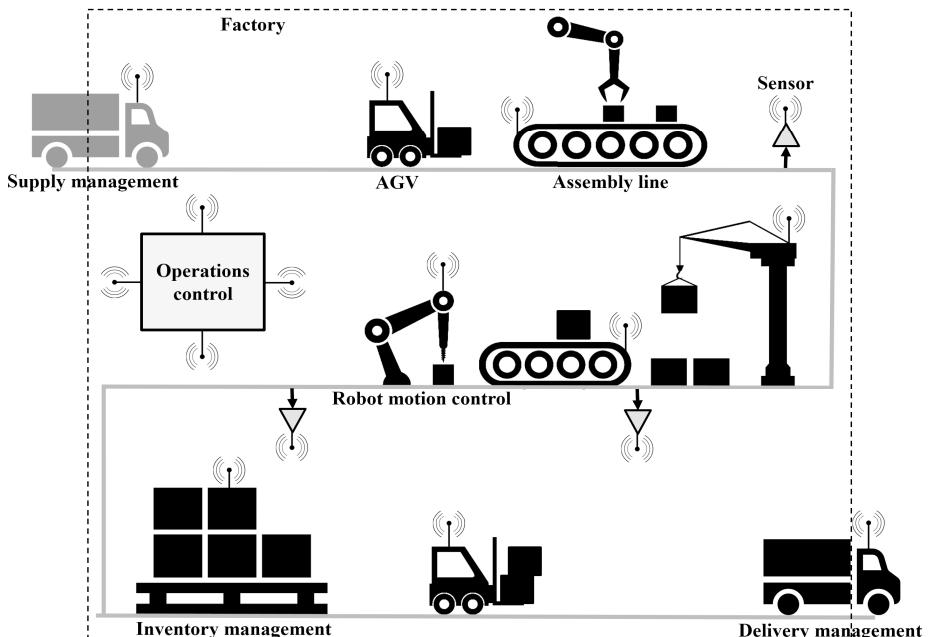


Fig. 4.3: Exemplary application areas of a 5G network in the factory of the future [71]

Figure 4.4 shows that such a networked factory of the future is dependent on the new features provided by 5G. It takes up the triangular structure shown in Figure 4.1 and links the basic ITU or 3GPP usage scenarios – eMBB, URLLC or CriC, and mMTC or MIoT – with possible use cases as shown in Figure 4.3 [71]. The arrangement illustrates that, depending on the application, very high data rates, very low delays, very high availability, very high connection density, or very high positioning accuracy must be guaranteed for the manufacturing area. Only a 5G network seems to be able to meet these requirements (see Table 4.3).

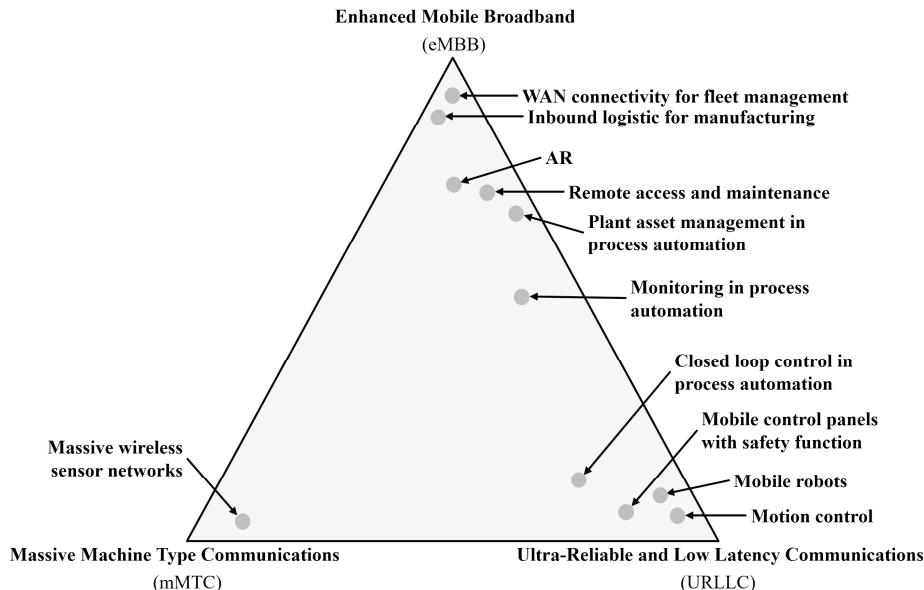


Fig. 4.4: 5G-ACIA use cases and ITU-R/3GPP usage categories [71]

The 5GAA (5G Automotive Association) was mainly concerned with the topic of automotive and 5G. It was formed to bring together cross-industry automotive, technology, and telecommunications companies to further develop future networked mobility in the context of 5G with a focus on V2X and to advance standardization accordingly [64]. Table 4.4 summarizes the use case groups developed by 5GAA and related use cases.

Tab. 4.4: V2X use cases [65]

Use case group	Use case
Safety	<ul style="list-style-type: none"> – Emergency braking – Intersection management assist – Collision warning – Lane change
Vehicle operations management	<ul style="list-style-type: none"> – Sensors monitoring – Software updates – Remote support
Convenience	<ul style="list-style-type: none"> – Infotainment – Assisted and cooperative navigation – Autonomous smart parking
Autonomous driving	<ul style="list-style-type: none"> – Control if autonomous driving is allowed – Tele-operation (potentially with AR support for a remote driver) – Handling of dynamic maps (update/download)
Platooning	<ul style="list-style-type: none"> – Collect and establish a platoon – Determine the position in the platoon – Dissolve a platoon – Manage distance within the platoon – Leave a platoon – Control of platoon in steady-state – Request passing through a platoon
Traffic efficiency and environmental friendliness	<ul style="list-style-type: none"> – Greenlight optimal speed advisory – Traffic jam information – Routing advice, e.g., smart routing
Society and community	<ul style="list-style-type: none"> – Emergency vehicle approaching – Traffic light priority – Patient monitoring – Crash report

A V2X application from Table 4.4 that is particularly interesting in practice is platooning, for which Figure 4.5 shows an example. Here, several vehicles, e.g., trucks, form a coherent group that moves together like a train. To maintain the distance between the vehicles, status information on speed, course, braking, acceleration, etc., must be exchanged. Also, other vehicles must be informed of the existence of a corresponding platoon in order not to interrupt or pass it. Information on the formation and cancellation of a platoon must also be exchanged. The advantages of platooning are that the distances between vehicles can be kept small, thus reducing overall fuel consumption due to slipstreaming, and only one driver is needed for the lead vehicle. As the distances between successive vehicles should be kept as short as possible, e.g., 1 m, but at the same time, the platoon should move at a speed of,

e.g., 100 km/h, short transit times for message exchange must be observed, e.g., 10 ms end-to-end. A solution based on the 5G network is, therefore, a good idea [26].

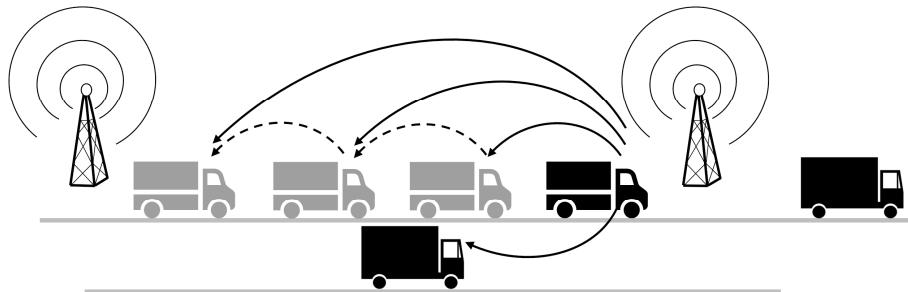


Fig. 4.5: Platooning with trucks

For all 5G applications mentioned at the beginning of this Section 4.2, there are documents from various organizations that describe use cases and requirements. As an example, the areas of manufacturing and industry 4.0, as well as automotive, were picked out above and deepened based on documents from 5G-ACIA, 5GAA, and 3GPP. In this context, we should mention the 3GPP study TR 22.806 [24], which in addition to use cases and requirements for the factories of the future, also deals with the usage areas public transport, energy supply, health, smart farming, and smart city.

Considering the comments on use cases and areas of application for 5G in Sections 4.1 and 4.2, we can conclude that the use case scenarios or categories identified by ITU-R in [128] and 3GPP in [27]

- eMBB (Enhanced Mobile Broadband),
- URLLC (Ultra-Reliable and Low Latency Communications) or CriC (Critical Communications), and
- mMTC (Massive Machine Type Communications) or MIoT (Massive Internet of Things)

including the corresponding general requirements summarised in Table 4.3 provide a good summary of the numerous conceivable and partly listed applications in the context of 5G.

4.3 5G Requirements

As discussed in Sections 4.1 and 4.2 above, the requirements for a 5G system should not primarily be based on the technical possibilities but on the conceivable use cases. Above, numerous use cases have been mentioned, and on this basis, the requirements derived from them have already been discussed. These we now expand and concretize in this section.

In Section 4.1, the IMT-2020 vision of ITU-R has already been mentioned. This is the first time an official standardization or regulatory organization has characterized a 5G target system with corresponding requirements in the Recommendation M.2083 [128]. They did it in continuation of the ITU-R specifications for the predecessor versions IMT-2000 (3G at 3GPP) and IMT-Advanced (4G at 3GPP). It should also be mentioned that, following ITU recommendations, only networks with systems from 3GPP Release 10 onwards with LTE-Advanced are officially described as 4th generation and thus IMT-Advanced [54]. In 3GPP terminology, LTE is already referred to as 4G from 3GPP Release 8 onwards.

ITU-R considers the following eight parameters to be the most important for an IMT-2020 or 5G system and has primarily specified maximum requirements for them, although not all criteria need to be fulfilled at the same time:

- Peak data rate: per user or UE up to 10 Gbit/s, under special conditions up to 20 Gbit/s
- User experienced data rate: per user or UE permanently 100 Mbit/s, at particular hotspots up to 1 Gbit/s
- Latency: for the RAN minimum of 1 ms
- Mobility: up to 500 km/h
- Connection Density: up to 10^6 UEs/km²
- Energy Efficiency: for the RAN 100 x better than IMT-Advanced, i.e., the same energy consumption at 100 times the performance. Here both the network and the end devices must be considered.
- Spectrum Efficiency: 3 x higher than IMT-Advanced
- Area Traffic Capacity: 10 Mbit/s/m².

Figure 4.6 shows these IMT-2020 requirement values compared to those of IMT-Advanced [128].

In addition to these key requirements, ITU-R also sees special demands on an IMT-2020 system in:

- Spectrum and Bandwidth Flexibility: use of different frequency ranges, even at higher frequencies, and larger channel bandwidths than IMT-Advanced
- Reliability: high service availability
- Resilience: correct continuation of operation during and after faults, e.g., after a power failure

- Security and Privacy: encryption and integrity protection for user data and signaling, protection of end-user privacy, protection of the network against fraud, hacking and denial of service attacks, etc.
- Operational Lifetime: e.g., battery life of more than 10 years for MTC end devices such as sensors [128].

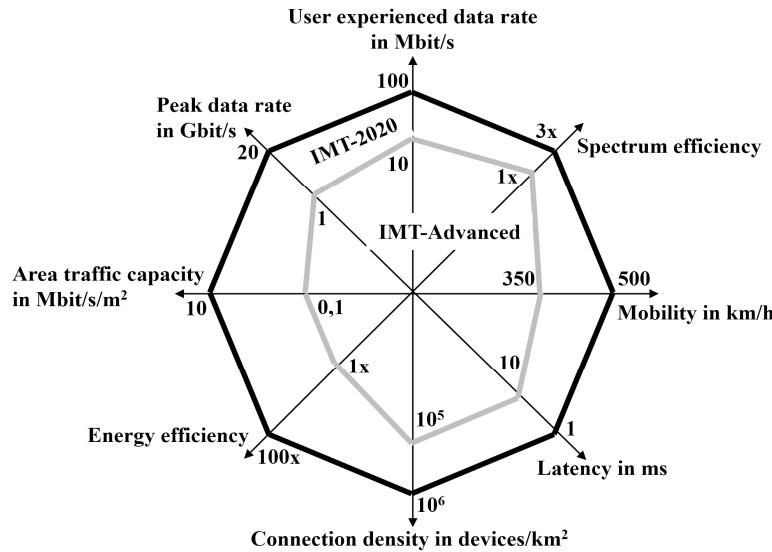


Fig. 4.6: Key capabilities of an IMT-2020 system (5G) in comparison with IMT-Advanced (4G) [128]

As already mentioned and clarified above, the various requirements formulated are all very ambitious. However, concerning the use cases or even the usage scenarios, it is advantageous that they never all have to be fulfilled simultaneously. Figure 4.7 shows this. For eMBB, latency and link density are of less importance; for URLLC, the latency requirement is dominant, while mMTC focuses on link density.

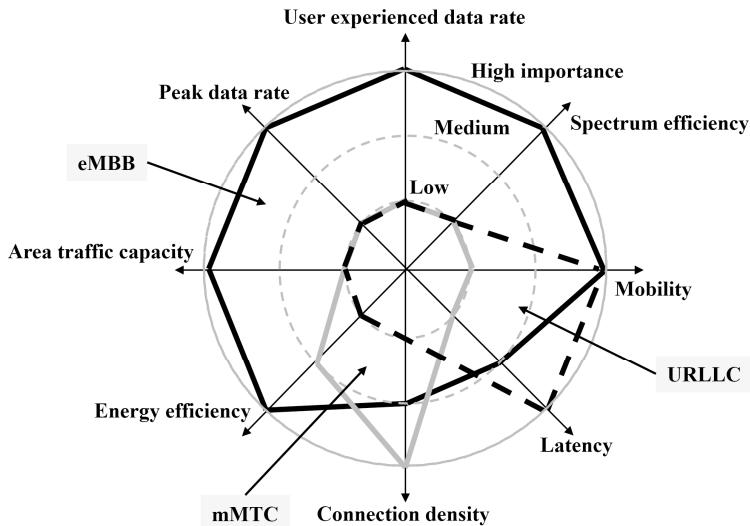


Fig. 4.7: IMT-2020 key capabilities in usage scenarios eMBB, URLLC, and mMTC [128]

ITU-R has described a 5G target system as outlined above. The concrete standardization, starting with the service requirements, has been and will be carried out at 3GPP, according to the current status for Releases 15, 16, and 17. Release 18 is in progress. In terms of requirements, the most crucial 3GPP specification is TS 22.261, which is available for each release with the corresponding extensions. Each of these documents organizes the requirements into five areas:

- Migration to 5G
- Basic capabilities
- Performance
- Security
- Charging aspects.

The following section describes the corresponding requirements, first for Release 15 and then for the follow-up Releases 16 and 17. This can only be an overview; details can be found in the respective standards.

In summary, a 5G system, according to [21], is characterized by:

- Support for multiple access network technologies
- Scalability and customizability
- Advanced KPI (Key Performance Indicator) values, e.g., on availability, latency, reliability, user experienced data rates, and area traffic capacity
- Flexibility and programmability through, e.g., network slicing, diverse mobility management, network function virtualization
- Resource efficiency regarding user plane and control plane

- Seamless mobility in both densely populated and heterogeneous environments
- Support for real-time and non-real-time multimedia services and applications with advanced QoE (Quality of Experience).

Tables 4.5 to 4.8 provide an overview of the service requirements for a 5G system according to 3GPP Release 15 [21].

Tab. 4.5: Service requirements regarding “Migration to 5G” in 3GPP Release 15 [21]

Requirements “Migration to 5G”

Support of most of the existing EPS services (Evolved Packet System, see Sections 2.5 and 2.6)

Interworking between 5G system regarding roaming

No seamless handover to 2G (GERAN) and 3G access networks (UTRAN), no access to a 5G core network via GERAN or UTRAN

Mobility management between 5G core und EPC (see Section 2.5)

Tab. 4.6: Service requirements regarding basic capabilities in 3GPP Release 15 [21]

Requirements “Basic capabilities”

Network slicing (see Section 8.4) to provide customized virtual networks with tailor-made functions for different requirements (see Sections 4.1 and 4.2)

Mobility management for UEs with stationary (e.g., sensor), nomadic (e.g., via wireline access), only local (e.g., in a factory) or network-wide use, service-specific uninterrupted mobility

Interoperable use of different access technologies: NR (New Radio), E-UTRA (Evolved-UTRA, LTE), non-3GPP (WLAN, wireline access). Selection of the most appropriate access network for the service. If necessary, simultaneous use of several access technologies by one UE

Resource efficiency in terms of control plane (signaling) and user plan (user data) despite different UEs (e.g., sensor, smartphone) and services (e.g., sensor status update, video streaming, cloud application)

Efficient handling of user data in the network even if the location of the user or application changes (e.g., service hosting changed due to latency requirements)

Efficient content delivery with flexible content caching, also for 3rd party providers (e.g., for frequently accessed video content)

Decoupled control of priorities and QoS, e.g., a high priority for rescue services and public safety; different priorities but the same QoS for airspace surveillance and UAV (Unmanned Aerial Vehicle)

Dynamic adaptation of policy control, e.g., for the prioritization of users or traffic or regarding QoS

Open network functions for 3rd party users via API, e.g., adapt network slice to customer requirements. Managing an application deployed on the network

Context-sensitive network by using existing information from sensors, the access networks, current traffic characteristics, etc.

Requirements “Basic capabilities”

Managing the subscription aspects of an IoT UE throughout its life cycle, e.g., when an IoT device changes location, network, owner, etc.

Energy efficiency, e.g., through energy-saving modes, an optimized operation for battery-powered terminals

Minimum service levels in different markets, e.g., give priority to health services in markets with low availability of electricity

Large ranges in sparsely populated areas, e.g., 100 km at 2 UEs/km²

A choice between all available access networks

Support of eV2X, e.g., for platooning

Shared use of a 5G access network (NG-RAN) by several network operators

Uniform access control for UEs

Tab. 4.7: Performance requirements for usage scenarios with high data rates and traffic densities [21]

Scenario	Data rate DL (Down Link)	Data rate UL (Up Link)	Area traffic capacity DL	Area traffic capacity UL	User density	Activity factor	Mobility
Urban macro	50 Mbit/s	25 Mbit/s	100 Gbit/s/km ²	50 Gbit/s/km ²	10000/km ²	20%	0-120 km/h
Rural macro	50 Mbit/s	25 Mbit/s	1 Gbit/s/km ²	500 Mbit/s/km ²	100/km ²	20%	0-120 km/h
Indoor hotspot	1 Gbit/s	500 Mbit/s	15 Tbit/s/km ²	2 Tbit/s/km ²	250000/km ²		0-5 km/h
Broadband access in a crowd	25 Mbit/s	50 Mbit/s	3,75 Tbit/s/km ²	7,5 Tbit/s/km ²	500000/km ²	30%	0-5 km/h
Dense urban	300Mbit/s	50 Mbit/s	750 Gbit/s/km ²	125 Gbit/s/km ²	25000/km ²	10%	0-60 km/h
Broadcast-like services	200Mbit/s per TV channel	500 kbit/s per user			15 TV channels		0-500 km/h
Train	50 Mbit/s	25 Mbit/s	15 Gbit/s/train	7,5 Gbit/s/train	1000/train	30%	0-500 km/h
Vehicle	50 Mbit/s	25 Mbit/s	100 Gbit/s/km ²	50 Gbit/s/km ²	4000/km ²	50%	0-250 km/h
Airplane	15 Mbit/s	7,5 Mbit/s	1,2 Gbit/s/airplane	600 Mbit/s/airplane	400/airplane	20%	0-1000 km/h

Tab. 4.8: Performance requirements for usage scenarios with low latency and high reliability [21]

Scenario	Max. end-to-end latency	Service availability [%]	Reliability [%]	Data rate	Traffic density	Connection density	Service area dimension
Industry automation	10 ms	99,99	99,99	10 Mbit/s	1 Tbit/s/km ²	100000/km ²	1000m x 1000m x 30m
Process automation – remote control	60 ms	99,999	99,999	1-100 Mbit/s	100 Gbit/s/km ²	1000/km ²	300m x 300m x 50m
Process automation – monitoring	60 ms	99,9	99,9	1 Mbit/s	10 Gbit/s/km ²	10000/km ²	300m x 300m x 50m
Electricity distribution – medium voltage	40 ms	99,9	99,9	10 Mbit/s	10 Gbit/s/km ²	1000/km ²	100 km along power line
Electricity distribution – high voltage	5 ms	99,999	99,999	10 Mbit/s	100 Gbit/s/km ²	1000/km ²	200 km along power line
Intelligent transport systems – infrastructure backhaul	30 ms	99,999	99,999	10 Mbit/s	10 Gbit/s/km ²	1000/km ²	2km along road

Besides, [21] specifies the following requirements:

- High accuracy positioning
- Security with features for authentication, authorization, identity management, regulatory compliance, and fraud protection
- Collecting information for charging aspects.

3GPP Release 15 defines a first 5G system with the requirements mentioned above. Based on this, Release 16 provides enhancements and completely new service requirements and features. Table 4.9 summarizes the most important service requirements for Release 16 [22; 185].

Tab. 4.9: Additional service requirements for a 5G system in 3GPP Release 16 [22; 185]**Requirements “Migration to 5G”**

Seamless handover for telephony service from NG-RAN to UTRAN with circuit switching (see Section 2.6)

Requirements “Basic capabilities”

IMS (See Section 2.2) as part of a network slice

Cross-network slice coordination

Satellite-based RANs

Mobility support across access network technologies. Mobility between the supported access networks, e.g., NG-RAN, WLAN, wireline broadband access, or wired broadband connections

Indirect and/or direct UE 5G network access. UE (e.g., a sensor in clothing, smart thermostatic valve, printer, smart flowerpot with remote irrigation) can be connected to the 5G system directly or via another UE acting as a relay station.

Independent radio-based network connection of 5G access nodes (self backhaul) via NG-RAN or E-UTRA (Evolved-Universal Terrestrial Radio Access, LTE-Advanced)

Flexible support of broadcast/multicast services (e.g., video streaming) in a specific geographical area

Simultaneous connection and service use over more than one 5G network

End-to-end QoS monitoring, especially for real-time services

Ethernet transport service by providing private LANs and virtual LANs in 5G network

Non-public 5G networks (e.g., by a company) in a defined geographical area, stand-alone, hosted, or implemented as a network slice

Position determination depending on the supported services (e.g., for emergency calls) with accuracies below 10 m

Communication services for Cyber Physical Systems (CPS) for intelligent control of physical processes with very high availability and often very short end-to-end latency, e.g., in the factory of the future (see Section 4.2), in the distribution and generation of electrical energy, in local rail transport

Messaging services for massive IoT communication with one-to-one, one-to-group, or one-to-all communication relationships with low delay and high availability

Performance requirements

AR or VR with audio-video synchronization

Radio-based backhaul infrastructure at the roadside, e.g., for connecting traffic light controls and traffic monitoring equipment to traffic control centers with high system availability (99,999%), low latency (30 ms end-to-end), and high connection density (1000/km²)

Positioning accuracy absolute up to 0.3 m horizontal and 2 m vertical when moving the UE at up to 60 km/h, among others

Security

For data stored in a cache, 5G-LAN

Enhancements regarding authentication, authorization (e.g., for IoT), fraud protection, data integrity and encryption

The contents of the above tables are so-called service requirements from the user's perspective. These are developed in an early phase of the standardization of a new 3GPP release and documented for each release in a TS 22.261 [21; 185; 22]. However, this does not mean that all the resulting necessary features and functions are standardized in the same release. An example is the satellite-based RAN mentioned in Table 4.9 for Release 16 (see Section 9.3). The result for this in Release 16 is only one study in the form of TR 22.822 [25]. The requirements for this in detail were the subject of Release 17, according to Table 4.10.

Tab. 4.10: Additional service requirements for a 5G system in 3GPP Release 17 [186; 243]

Requirements “Basic capabilities”

Service continuity, when the remote UE changes from a direct network connection to an indirect network connection and vice-versa

Relay UEs that support multiple access types like 5G RAT, WLAN access, fixed broadband access

RAN Slicing

Service continuity between 5G terrestrial access and 5G satellite access networks

Roaming of UE supporting both satellite access and terrestrial access

Satellite based RANs (see Section 9.3)

Meshed connectivity between satellites

Efficient bulk operations for IoT with up to 1000000 connections/km²

Broadcasting/multicasting per satellite on very large to global coverages

Use of relay UEs (also multiple hops)

Authorization of a UE as a relay UE

Relay UEs with satellite access

5G LAN-type service using an indirect network connection

Roaming control enhancements

Minimization of service interruption in the event of a disaster (e.g., a fire) by obtaining connectivity service (e.g., voice call) of another PLMN of this area, roaming of the UE

Controlling and providing services (e.g., video) for UAVs (e.g., drone), incl. operation of the on-board radio interfaces

Video, imaging, and audio for professional applications such as audiovisual productions (e.g., in television and radio studios, at sports events or music festivals) with wireless devices networked via 5G (e.g., microphone, monitoring system, camera)

Communication services for critical medical applications (e.g., remote diagnosis, monitoring or surgery, AR)

Performance requirements

Service provision via satellite access with an end-to-end delay of up to 285 ms for GEO (Geostationary Earth Orbit), 95 ms for MEO (Medium Earth Orbit), and 35 ms for LEO (Low Earth Orbit) satellites

Very high availability of > 99.9999% for IoT traffic with up to 1000 UEs/km²

High data rates and low end-to-end latency: ≤ 10 Gbit/s, 10 ms for VR; ≤ 1 Gbit/s, 10 ms for Gaming

For indirect UE-5G network connection via relay UE: ≤ 1 Gbit/s, 10 ms with 50 UEs/home; ≤ 5 Mbit/s, 50 ms to 1 s with 10000 UEs/factory, etc.

Security

Secure mechanism to protect relayed data from being intercepted by a relay UE

The final features of a release in terms of technical implementation are summarized in a Technical Report (TR) at the end of the standardization work for a release – when all technical specifications (TS) are available: for Release 15 in TR 21.915 [19], for Release 16 in TR 21.916 [20], and for Release 17 in TR 21.917 [243]. For Release 18, which is still in the working process, it was, of course, not yet available in September 2023.

5 5G Standardization and Regulation

As mentioned in Chapter 4, the ITU has specified a 5G target system under the IMT-2020 designation. 3GPP did and does the actual standardization for 5G. This was already evident when the requirements were formulated in Section 4.3. The simple reason for this is that 3GPP was founded by the relevant European, Asian, and North American standardization organizations specifically for global standardization on mobile networks, starting with 3G, and 3GPP simply continues its overall mission for 5G.

The following sections are discussing the possible frequencies for 5G first. Then the standardization for 3GPP in general and specifically for 5G is explained. Finally, the national implementation, i.e., the country-specific regulation of 5G, is examined more closely.

5.1 Frequencies

The frequency spectrum available for radio communications is a scarce resource. Nevertheless, it would be desirable for 5G to have globally uniform areas with sufficient bandwidth for the desired high bit rates, which are advantageously not such high-frequency areas due to the geographical coverage of a radio cell. A possible implementation would keep the number of base stations required and the complexity of the hardware, especially the highly integrated circuits for the radio interfaces, relatively low. However, these coherent and globally available frequency ranges do not exist. Therefore, a relatively large number of country-specific frequency bands must be used for 5G. Besides, it is necessary to include higher frequency ranges not previously considered for mobile radio [88].

The WRCs (World Radiocommunication Conference), organized by the ITU, specify the frequency bands for the various radio services. WRC-15 in 2015 limited the frequency ranges for mobile communications to below 6 GHz, but WRC-19 in 2019 has also opened up areas above 24 GHz [176].

Based on the WRC definitions, 3GPP has identified and mapped frequency ranges that can be used in 5G RANs, with country-specific availability based on national regulation. 3GPP distinguishes in Releases 15 and 16 [47; 187] between a lower and a higher frequency range:

- FR1 (Frequency Range): 410 – 7125 MHz
- FR2: 24,25 – 52,6 GHz.

FR1 covers, on the one hand, the previous 2G to 4G ranges. On the other hand, comparatively low-frequency ranges are included, which are very well suited for MTC or IoT since long ranges and good penetration of radio obstacles such as house walls

are given. FR2 specifies completely new frequency spectra for mobile radio in the range of cm waves up to 30 GHz and mm waves above 30 GHz, whereby it is usual to speak of mm waves in this entire frequency range. However, this results in only relatively short radio ranges and significant impairments due to obstacles (e.g., water vapor, fog, rain, leaves, also people) in the radio path [127]. Accordingly, radio transmission and antenna technology must become more complex. Also, an application is locally limited. To be able to provide additional usable frequency ranges, Release 17 extends the FR2 to include an even higher frequency range from 52.6 to 71 GHz despite the disadvantages mentioned above.

Table 5.1 shows in detail the FR1 frequency bands according to 3GPP for the 5G radio transmission technology NR (New Radio) for the uplink (UL) or downlink (DL) direction, i.e., from UE to BS or vice versa. The table also indicates the directional separation method to be used, with FDD (Frequency Division Duplexing) at different frequencies in UL and DL, with TDD (Time Division Duplexing) at different times. SUL (Supplementary Uplink) or SDL (Supplementary Downlink) designates frequency ranges exclusively for UL or DL to enable an asymmetrical and, thus, higher bit rate in combination with FDD or TDD. FR1 operates with channel bandwidths between 5 and 100 MHz [47].

Table 5.2 provides a corresponding overview of the new frequency bands defined by 3GPP for 5G in the so-called mm range. The bandwidths here are between 50 and 400 MHz [47]. Interestingly, FR2 differs from the results of WRC-19. WRC-19 has identified the ranges 24.25-27.5 GHz, 37-43.5 GHz, 45.5-47 GHz, 47.2-48.2, and 66-71 GHz [177].

The spectra mentioned above are licensed, i.e., partial frequency bands are allocated by a regulatory authority for a specific geographical region exclusively to exactly one network operator, e.g., in an auction. In this case, network planning is relatively simple because there is no interference from the radio channels of other network operators. In 3GPP Release 15, only licensed frequency ranges are used for NR (for LTE and NB-IoT in 5G, unlicensed frequency ranges are already available). As of Release 16, however, unlicensed spectra for NR such as WLAN (2.4, 5, and 6 GHz) or LPWAN (Low Power Wide Area Network, above 433, 863, or 902 MHz) are also considered.

Tab. 5.1: FR1 frequency bands for 5G according to 3GPP [47; 187; 188]

NR frequency band	UL (Up Link)	DL (Down Link)	Duplex mode
n1	1920 MHz – 1980 MHz	2110 MHz – 2170 MHz	FDD
n2	1850 MHz – 1910 MHz	1930 MHz – 1990 MHz	FDD
n3	1710 MHz – 1785 MHz	1805 MHz – 1880 MHz	FDD
n5	824 MHz – 849 MHz	869 MHz – 894 MHz	FDD
n7	2500 MHz – 2570 MHz	2620 MHz – 2690 MHz	FDD
n8	880 MHz – 915 MHz	925 MHz – 960 MHz	FDD
n12	699 MHz – 716 MHz	729 MHz – 746 MHz	FDD
n20	832 MHz – 862 MHz	791 MHz – 821 MHz	FDD
n25	1850 MHz – 1915 MHz	1930 MHz – 1995 MHz	FDD
n28	703 MHz – 748 MHz	758 MHz – 803 MHz	FDD
n34	2010 MHz – 2025 MHz	2010 MHz – 2025 MHz	TDD
n38	2570 MHz – 2620 MHz	2570 MHz – 2620 MHz	TDD
n39	1880 MHz – 1920 MHz	1880 MHz – 1920 MHz	TDD
n40	2300 MHz – 2400 MHz	2300 MHz – 2400 MHz	TDD
n41	2496 MHz – 2690 MHz	2496 MHz – 2690 MHz	TDD
n50	1432 MHz – 1517 MHz	1432 MHz – 1517 MHz	TDD
n51	1427 MHz – 1432 MHz	1427 MHz – 1432 MHz	TDD
n66	1710 MHz – 1780 MHz	2110 MHz – 2200 MHz	FDD
n70	1695 MHz – 1710 MHz	1995 MHz – 2020 MHz	FDD
n71	663 MHz – 698 MHz	617 MHz – 652 MHz	FDD
n74	1427 MHz – 1470 MHz	1475 MHz – 1518 MHz	FDD
n75	–	1432 MHz – 1517 MHz	SDL
n76	–	1427 MHz – 1432 MHz	SDL
n77	3300 MHz – 4200 MHz	3300 MHz – 4200 MHz	TDD
n78	3300 MHz – 3800 MHz	3300 MHz – 3800 MHz	TDD
n79	4400 MHz – 5000 MHz	4400 MHz – 5000 MHz	TDD
n80	1710 MHz – 1785 MHz	–	SUL
n81	880 MHz – 915 MHz	–	SUL
n82	832 MHz – 862 MHz	–	SUL
n83	703 MHz – 748 MHz	–	SUL
n84	1920 MHz – 1980 MHz	–	SUL
n86	1710 MHz – 1780 MHz	–	SUL
From Release 16:			
n89	824 MHz – 849 MHz	–	SUL
n90	2496 MHz – 2690 MHz	2496 MHz – 2690 MHz	TDD
n91	832 MHz – 862 MHz	1427 MHz – 1432 MHz	FDD

NR frequency band	UL (Up Link)	DL (Down Link)	Duplex mode
n92	832 MHz – 862 MHz	1432 MHz – 1517 MHz	FDD
n93	880 MHz – 915 MHz	1427 MHz – 1432 MHz	FDD
n94	880 MHz – 915 MHz	1432 MHz – 1517 MHz	FDD
n95 (only in China)	2010 MHz – 2025 MHz	–	SUL
n96 (a.o. in USA)	5925 MHz – 7125 MHz	5925 MHz – 7125 MHz	TDD
From Release 17:			
n97	2300 MHz – 2400 MHz	–	SUL
n98	1880 MHz – 1920 MHz	–	SUL
n99	1626,5 MHz – 1660,5 MHz	–	SUL
n100	874,4 MHz – 880 MHz	919,4 MHz – 925 MHz	FDD
n101	1900 MHz – 1910 MHz	1900 MHz – 1910 MHz	TDD
n102	5925 MHz – 6425 MHz	5925 MHz – 6425 MHz	TDD
n104	6425 MHz – 7125 MHz	6425 MHz – 7125 MHz	TDD

Tab. 5.2: FR2 frequency bands for 5G according to 3GPP [47; 187; 188]

NR frequency band	UL	DL	Duplex mode
n257	26500 MHz – 29500 MHz	26500 MHz – 29500 MHz	TDD
n258	24250 MHz – 27500 MHz	24250 MHz – 27500 MHz	TDD
n260	37000 MHz – 40000 MHz	37000 MHz – 40000 MHz	TDD
n261	27500 MHz – 28350 MHz	27500 MHz – 28350 MHz	TDD
From Release 16:			
n259	39500 MHz – 43500 MHz	39500 MHz – 43500 MHz	TDD
From Release 17:			
n262	47200 MHz – 48200 MHz	47200 MHz – 48200 MHz	TDD
n263	57000 MHz – 71000 MHz	57000 MHz – 71000 MHz	TDD

In addition to the above comments on the frequency ranges specified by 3GPP for 5G, the advantages and disadvantages of the different frequency ranges are explained below [196].

“Depending upon frequency bands, the technology will perform differently, and some bands will be better suited for certain use cases than others.

For instance, lower frequency bands, such as those below 2 GHz, are an excellent fit for coverage and mobility and are valuable for high aggregation of low bandwidth users, such as interactive communications and massive Machine Type Communications (mMTC). The low-band spectrum is also well suited for indoor penetration. In terms of capacity, some 5G use cases will rely on significantly higher

peak data rates for faster connections and low latency, and this will require wider channels than are available in the lower bands.

Higher frequency bands, such as those in the millimeter waves (mmW), are optimal for short range, low latency, and very high capacity transmissions for enhanced mobile broadband (eMBB), but with a more limited range and with limited indoor penetration.

Mid-band spectrum offers a balance of these capabilities, complementary to mmW in urban and suburban settings, and extending the availability of 5G beyond densely populated areas. Mid-band deployments typically use a smaller number of macro base stations – in contrast to the larger number of small cells required to support mmW 5G deployments.

Each spectrum range has specific characteristics, as previously explained, that makes it more suitable for certain deployment scenarios. While the low range of spectrum has very good propagation aspects that make it feasible for large area coverage, low-band has limited capacity due to the lack of available spectrum and component design considerations. The mid-range of spectrum provides a type of coverage more feasible for urban deployment due to increased capacity. The high-range of spectrum is more limited in coverage but could provide very high capacity due to the amount of unused spectrum and wider channelization available at these frequencies.” [196]

5.2 Standardization

As mentioned above, two global standardization organizations were and are active in 5G. This is the ITU, on the one hand, the ITU-R SG5 (Study Group 5) WP5D (Working Party 5D: IMT-Systems) for radio technology, on the other hand, the ITU-T SG13 (Future networks, with focus on IMT-2020, cloud computing, and trusted network infrastructures) for the network aspects. A 5G target system was specified by the ITU-R under the keyword IMT-2020. The 5G standardization at 3GPP as a basis for system and network implementations was and is much more comprehensive and detailed. Because of the importance of 3GPP, the following section will take a closer look at the 3GPP organization and working methods to understand and follow the development of 5G releases.

According to Figure 5.1, 3GPP is organized in three areas, each with working groups. The working groups document their results: interim results and studies in the form of non-binding Technical Reports (TR), the actual binding standards as Technical Specifications (TS). Only functions and protocols are described, not how to implement them. A related set of TS and TR documents represents a system version, a release.

Project Coordination Group (PCG)		
TSG (Technical Specification Group) CT Core • Network & Terminals	TSG RAN • Radio Access Network	TSG SA • Service & System Aspects
CT WG1 • User Equipment to Core Network protocols	RAN WG1 • Radio Layer 1 (Physical layer)	SA WG1 • Services
CT WG3 • Interworking with External Networks & Policy and Charging Control	RAN WG2 • Radio Layer 2 and Radio Layer 3 Radio Resource Control	SA WG2 • System Architecture and Services
CT WG4 • Core Network Protocols	RAN WG3 • UTRAN/E-UTRAN/NG-RAN architecture and related network interfaces	SA WG3 • Security and Privacy
CT WG6 • Smart Card Application Aspects	RAN WG4 • Radio Performance and Protocol Aspects	SA WG4 • Multimedia Codecs, Systems and Services
	RAN WG5 • Mobile terminal conformance testing	SA WG5 • Management, Orchestration and Charging
	RAN AH1 • ITU-R Ad Hoc	SA WG6 • Application Enablement and Critical Communication Applications

Fig. 5.1: 3GPP organization for mobile communications and 5G standardization [55]

The standardization work is divided into phases (Stage 1 to 3) per release. Stage 1 describes the services to be provided by the target system from the user's perspective. In Stage 2, the necessary network functions and their interaction are worked out according to the service requirements. Finally, Stage 3 specifies the concrete switching functions and protocols required for the services defined in Stage 1. For the verification of the results in Stage 3, program code is partly developed, e.g., in the form of ASN.1 code. This is then part of a specially designated ASN.1 phase, which completes the release standardization [57].

In line with this common approach to 3GPP, Release 15 (5G Phase 1) [59] was standardized for a first complete 5G system by mid-2019. As an intermediate step, referred to as "early drop" in Figure 5.2, a so-called non-standalone 5G system (NSA) was specified for the first 5G implementations, which connects NR-5G access network technology with correspondingly high bit rates to the 4G core network with EPC (see Section 2.5). In this case, we still have a 4G network, but with 5G-RAN connected and, therefore, higher data rates.

Release 16 (5G Phase 2) [60] was developed by the end of 2020. The aim was that a 5G system based on the Release 16 standards should comply with the IMT-2020 target system defined by the ITU. Finally, Release 17 followed in the third quarter of 2022 [61]. The time sequences and relationships for all three of these releases are

shown in Figure 5.2 [57]. The requirements developed in Stage 1 have already been discussed in Section 4.3. In addition to the above comments on the standardization procedure for 3GPP, it should be added concerning Figure 5.2 that the final RF (Radio Frequency) and performance specifications for both base stations and terminal equipment are developed in the RAN4 phase.

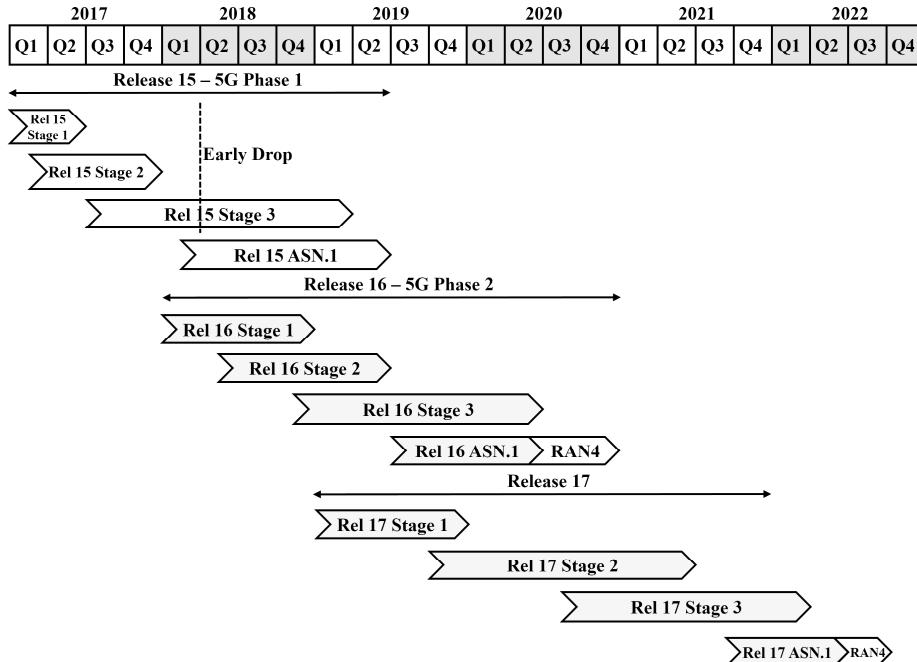


Fig. 5.2: Standardisation procedures for 3GPP for Releases 15, 16, and 17 [57]

5.3 Regulation

As mentioned in Section 5.1 above, although 3GPP specifies possible frequency ranges for use in 5G by default, the actual making available of licensed frequency ranges to network operators is done per country by the national regulatory authorities. In the German-speaking area, these are for

- Germany the Bundesnetzagentur (BNetzA) [81], for
- Switzerland the Eidgenössische Kommunikationskommission (ComCom) with the cooperation of the Bundesamt für Kommunikation (BAKOM) [74] and for
- Austria, the Telekom-Control-Kommission (TKK) within the Rundfunk und Telekom Regulierungs-GmbH (RTR) [157].

In Germany, a frequency auction for 5G took place in 2019. Frequencies in the ranges around 2 GHz and above 3.4 GHz were auctioned off and thus licensed. Table 5.3 shows the results [82; 155]. The permanent use of the auctioned frequencies is subject to conditions set by the BNetzA. Among other things, the following supply requirements apply to every network operator:

- By the end of 2022, at least 98% of all households, all federal motorways, part of the federal roads, and highly frequented railways with at least 100 Mbit/s
- By the end of 2024, all other national roads with 100 Mbit/s, state roads, seaports, and major waterways and other railways with 50 Mbit/s
- 1000 5G base stations by the end of 2022, plus an additional 500 base stations with at least 100 Mbit/s in previously uncovered areas.
- Weaker supply conditions apply to new entrant network operators.

If these expectations are not met, the frequencies fall back to the BNetzA [83].

According to [83], in addition to the nationwide frequency usage rights listed in Table 5.3, the BNetzA provides additional frequencies in the 3.7 GHz to 3.8 GHz and 26 GHz ranges for local allocation and use. In particular, the spectral range above 3.7 GHz is intended for companies and organizations with their own, geographically limited 5G network.

Tab. 5.3: Results of the 5G frequency auction of June 2019 in Germany [82; 155]

Frequency ranges	Duplex mode	Bandwidth	Number of network operators
1,92 – 1,98 GHz, 2,11 – 2,17 GHz	FDD	2 x 60 MHz	4
3,4 – 3,7 GHz	TDD	300 MHz	4

In Switzerland, the first 5G frequency auction was already held in February 2019. The results are given in Table 5.4 [75; 76].

The Swiss regulatory authority also imposes conditions of use [76]:

- With licensed frequencies in the 700 MHz FDD range, at least 50% of all households must be covered by the end of 2024.
- For frequency ranges above 700 MHz, this applies to 25% of all households.

In the event of non-performance, the rights of use may be withdrawn without compensation.

Tab. 5.4: Results of the 5G frequency auction of February 2019 in Switzerland [75; 76]

Frequency ranges	Duplex mode	Bandwidth	Number of network operators
703 – 733 MHz, 758 – 788 MHz	FDD	2 x 30 MHz	3
738 – 753 MHz	SDL	10 MHz	1
1,427 – 1,517 GHz	SDL	75 MHz	3
3,5 – 3,8 GHz	TDD	300 MHz	3

The first auction of the 5G spectrum in Austria was completed in March 2019. Table 5.5 shows the result. Of particular interest is the fact that in addition to a federal allocation to three network operators, there was also a regionally limited allocation to four other providers [171].

[172] shows the coverage obligations in Austria's 5G networks resulting from the licensed frequencies:

- For each region, an area-dependent minimum number of base station sites is defined for the end of 2020 and mid-2022.
- The number of base stations required depends not only on the region but also on the bandwidth purchased: the more bandwidth, the more locations.

If the supply obligations are not fulfilled, contractual penalties are incurred.

Tab. 5.5: Results of the 5G frequency auction of March 2019 in Austria [171; 155]

Frequency ranges	Duplex mode	Bandwidth	Number of network operators
3,4 – 3,6 GHz, individual areas	TDD	80 MHz	4
3,6 – 3,8 GHz, nationwide	TDD	325 MHz	3

Looking at the EU as a whole [197], countries use low (700 – 900 MHz), mid (1,5 – 2,6 GHz, 3,4 – 3,8 GHz), and in some cases, high-frequency ranges (26 GHz band). The mid-band spectrum is defined as the baseline capacity layer in favor of flexibility for many use cases with higher throughputs, wider spectrum, and potential re-farming of the LTE spectrum. The 3.4-3.8 GHz band is the primary band in Europe with early availability.

The high-band spectrum is known as the extreme capacity layer with large amounts of spectrum potentially available for very high capacity, very high data rates but limited coverage. The 26 GHz band (24.25 – 27.5 GHz) is the pioneer high

band for 5G in Europe. Italy was the first EU member state to allow spectrum use for 5G in all pioneer bands, incl. 26 GHz band (700 MHz band, 3,4 – 3,8 GHz, 26 GHz band) in October 2018. Finland followed in June 2020 [155].

The view beyond Europe leads to the USA. Here, the regulatory authority, the FCC (Federal Communications Commission) [194], has organized several frequency auctions for 5G, according to Table 5.6 [193]. Noticeably, until the beginning of 2020, the focus was mainly on spectra with millimeter waves up to 48 GHz. Then, in the fall of 2020, the FCC began allocating the mid-range 5G frequencies starting at 3.55 GHz, preferred in Europe.

Tab. 5.6: 5G frequency auctions in the USA [193; 206]

Auction	Frequency range	Bandwidth	Date
1000 (Broadcast)	575 – 698 MHz	70 MHz	2016
101	27,5 – 28,35 GHz	850 MHz	January 2019
102	24,25 – 24,45 GHz 24,75 – 25,25 GHz	200 MHz 500 MHz	May 2019
Order (CBRS)	2,496 – 2,69 GHz	116,5 MHz	July 2019
103	37,6 – 38,6 GHz 38,6 – 40 GHz 47,2 – 48,2 GHz	1 GHz 1,4 GHz 1 GHz	March 2020 March 2020 March 2020
105	3,55 – 3,65 GHz	70 MHz	September 2020
107	3,7 – 3,98 GHz	280	2021
110	3,45 – 3,55 GHz	100 MHz	2022

In the USA, the frequency range from 3.55 to 3.7 GHz is used as the so-called CBRS band (Citizens Broadband Radio Service). In 2015, FCC adopted rules to accommodate shared federal (e.g., US navy radar operators) and non-federal use of this band. Access and operations will be managed by an automated frequency coordinator, known as a Spectrum Access System (SAS). In the CBRS frequency range, network operators (non-federal) can offer 5G mobile services without having to acquire special frequency licenses. However, they require Priority Access Licenses (PALs), which are awarded on a county-by-county basis through competitive bidding within the 3,55 – 3,65 GHz band (see auction 105 in Table 5.6) [193]. Each PAL is defined as a renewable license to use a 10 MHz channel in a single specific closed area (county) for ten years. Up to seven PALs can be allocated per region. The way the CBRS band is used for 5G enables efficient spectrum usage by introducing small cells and spectrum sharing [199].

In July 2019, the FCC adopted an order remaking the 2.6 GHz band, previously allocated to the Educational Broadband Service (EBS). The order eliminated re-

strictions on what types of entities can hold licenses in that band. The new licenses are offered on a county-by-county basis as overlay licenses that give the licensees the right to operate anywhere there is not an incumbent educational or Tribal licensee in place [200].

In the meantime, frequencies in the 600 MHz band are also being used for 5G according to Table 5.6. They were already returned by broadcasters in 2016 for compensation (reverse auction) and then auctioned again (forward auction) [206; 207]. This low-frequency band, like the 700 MHz band used in Europe and China, enables radio cells with a large coverage area of several kilometers.

The use of the so-called C-band (3,4 – 4,2 GHz), previously used for satellite services, as the primary 5G band is interesting both in terms of 5G frequency allocation in the USA and Europe. The C-band offers coverage of continental zones and was assigned to Fixed Satellite Services (FSS). Frequency allocation for the downlink in the US is from 3,7 GHz to 4,2 GHz and in Europe from 3,4 GHz to 4,2 GHz. The C band is ideal for supporting telecommunications and broadcasting services in rural and marine areas, where terrestrial infrastructure is sparse or does not exist. Another benefit of the C band is its low susceptibility to rain fade, which qualifies it for stable links in tropical areas. Nevertheless, in March 2020, the FCC released the final decision on repurposing the C-band spectrum. The lower 280 MHz of the 3,7 – 4,2 GHz range shall be cleared no later than December 2025. Satellite operators involved in the process need to migrate their C-band services to 4,0 – 4,2 GHz, for which they will be reimbursed for relocation costs [198].

As we have already seen, C-band is essential for 5G in Europe. But the situation is not comparable to that in the USA. In Europe, the use of the C band has been declining for some time. The shift is towards fiber-based transport and satellite services in Ku and Ka band. In addition, the C band is rarely used for satellite television in Europe. Moreover, frequency allocations for 5G in Europe are only in the 3,4 – 3,8 GHz range, which does not put significant pressure on satellite services in 3,7 – 4,2 GHz [198].

Finally, for the regulatory section, Table 5.7 provides an overview of countries and regions where the 5G rollout is already well advanced. As of March 2023, the table shows how much spectrum in which frequency bands have already been allocated to network operators by the regulatory authorities and how far the rollout has progressed. Average values for all 27 member states were calculated for the EU concerning the allocated spectrum [206].

Tab. 5.7: Allocated frequency spectrum and 5G network rollout in international comparison [206; 208]

Country/ Region	Frequency spectrum	Assigned spec- trum bands	Number of base stations	Base stations per 100000 inhabitants
Germany	700 MHz (703-788) 3,6 GHz (3,4-3,8) 26 GHz (24,25-27,5)	60 MHz 400 MHz —	65905	79
EU	700 MHz (703-788) 3,6 GHz (3,4-3,8) 26 GHz (24,25-27,5)	44 MHz 335 MHz 291 MHz	256074	57
USA	600 MHz (575-698) 2,6 GHz (2,496-2,69) 3,45 – 3,55 GHz 3,5 – 3,7 GHz 3,7 – 3,98 GHz 24 GHz (24,25-25,25) 28 GHz (27,5-28,35) 39 GHz (37,6-40) 47 GHz (47,2-48,2)	70 MHz 605 MHz in total 4950 MHz in total —	100000	30
Japan	3,6 – 4,2 GHz 4,4 – 4,9 GHz 28 GHz (27-29,5)	880 MHz in total 1600 MHz	50000	40
South Korea	3,6 GHz (3,42-3,8) 28 GHz (26,5-28,9)	380 MHz 2400 MHz	215000	415
China	700 MHz (703-788) 2,6 GHz (2,51-2,675) 3,3 – 3,6 GHz 4,8 – 5,0 GHz 24 GHz etc.	80 MHz 660 MHz in total —	1850000	132

6 5G Networks at a Glance

6.1 Design Principles

As we have seen in Chapter 4, a 5G network poses significant network design challenges due to the wide range of use cases and categories with extreme demands on functionality, flexibility, and performance to be supported. It is, therefore, not surprising that the technologies discussed in Chapter 3 as essential building blocks of current and future modern networks are also very relevant for the design of 5G networks. These are

- NFV (Network Functions Virtualisation) with the orchestration of the network functions (see Section 3.1),
- SDN (Software Defined Networking) (see Section 3.2),
- MEC (Multi-access Edge Computing) (see Section 3.1), and
- C-RAN (Cloud-RAN or also Centralized-RAN) (see Section 3.1).

Besides, a basic design principle of 4G networks is maintained: the All-IP network. Despite the use of state-of-the-art network technology, IP remains the primary protocol for 5G.

As already mentioned in Chapter 4, the very different and sometimes extreme requirements for eMBB (Enhanced Mobile Broadband), URLLC (Ultra-Reliable and Low Latency Communications), and/or mMTC (Massive Machine Type Communications) cannot be met by a monolithic 5G system at all. One monolithic system cannot cover the entire range of requirements. The solution to this problem is to have requirement-specific subsystems within an overall 5G system, whose respective functionality is assembled from modular network functions. The design principle for this is modularization: there are modules for Access Network (AN) and Core Network (CN) functions for the Control Plane (CP) with the signaling and control protocols and the User Plane (UP) for the user data. These functional modules are put together and combined according to requirements. These are relatively fine-grained network functions (NF) that are provided in a repository and can be called via APIs. This concept is called Service Based Architecture (SBA). But so far, it is only applied to the CN.

The NFs are implemented via NFV. Their combination into service chains and the formation of subsystems within the framework of network slices are done via NFV and SDN. The latter two techniques, which are existential for 5G, are summarized under the keyword Network Softwarization.

Modularization and network softwarization provide not only an excellent technical basis for the provision of various services with diverging requirements but also for the use of a 5G network by several tenants. These could be, for example, a mobile network provider and operators of their virtual subnetworks from the energy, auto-

motive, or health industry. Let us spin this idea even further. Network or IT infrastructure providers and telecommunications network operators can also participate in the same 5G network. This multi-tenant capability is guaranteed not only by the design principles of modularization and network softwarization with network slicing mentioned above, but also by shared computer, storage, and network hardware. These results – wherever possible – in the use of standard server hardware and the execution of the software representing the network functions as cloud applications (cloudification). The application of these design principles also leads to a minimization of the system and operating costs.

The multi-tenant capability mentioned above also brings with it the desire for openness for 3rd party providers, coupled with corresponding APIs and the possible use of MEC (e.g., for optimized video delivery, local content caching, car-to-x communication, or an IoT gateway).

The very different requirements for radio technology with very high bit rates in eMBB or long ranges and penetration of obstacles such as building walls in mMTC mean that we have to support heterogeneous RAN technology in different frequency bands (< 6 GHz or mm waves) with varying sizes of a cell (a few meters to hundreds of kilometers).

With the aim of a convergent network, interoperable use of different access technologies must be possible, not only NR and E-UTRA, but also WLAN and, above all, high bit-rate wireline access (non-3GPP).

This, in turn, means that the core network with its functions should be decoupled and thus independent of the access network technology used.

Also, the heterogeneous, partly service-specific access networks require that a UE must be able to be connected to different access networks at the same time, i.e., flexible UE connectivity must be supported.

Also, backward (interworking with 4G) and upward compatibility (> Release 16) must and has been ensured concerning acceptance, costs, and early market availability of a 5G system [133; 88].

In summary, the following design principles and key technologies are the basis of a 5G system:

- All-IP network
- Modularization with SBA
- Network softwarization with NFV, SDN, and network slicing
- Multi-tenant capability
- Cloudification, including C-RAN and MEC
- Openness for 3rd party providers
- Heterogeneous RAN technology
- Various radio and wired access network technologies
- Core network decoupled from access network technology
- Flexible terminal device connectivity
- Downward and upward compatibility.

6.2 Features and Functions

Based on the requirements for a 5G system outlined in Section 4.3, the features and functions were defined. Table 6.1 provides an overview of Release 15 [19]. According to the 3GPP standardization shown in Figure 5.2, this is the 5G system – phase 1. 3GPP divided this phase 1 into intermediate steps with implemented intermediate results, the early drop, the main drop, and the late drop. The latter characterizes the complete Release 15.

Tab. 6.1: Features and functionalities of 3GPP Release 15 [59; 19]

Features
5G system – phase 1
NSA architecture (Non-Standalone) – early drop
SA architecture (Standalone) – main drop
5G access network
<ul style="list-style-type: none"> – NR, for FR1 and/or FR2 – gNB can be divided into gNB-CU (Control Unit) and gNB-DU (Distributed Unit) – Split CU into CU-UP and CU-CP – Dual Connectivity – Coexistence with LTE
5G core network
<ul style="list-style-type: none"> – Service Based Architecture (SBA) – Network Slicing – Local hosting of services and edge computing – Uniform access control – Support of 3GPP- und Non-3GPP access networks – Framework for policy control and QoS support – Make network functions available to 3rd party providers – IMS optional
Security model for NSA and SA
Enhancements for Mission Critical communication (MC; low latency, high availability) with 5G- or 4G technology (EPC, LTE)
Enhanced performance for MTC and IoT applications
Vehicle-to-Everything (V2X) – phase 2
<ul style="list-style-type: none"> – Platooning – Integration of information from remote sensors (e.g., in a vehicle) into the own view of a pedestrian or other vehicle – Autonomous driving – Driving with remote control
WLAN for
<ul style="list-style-type: none"> – Proximity-based Services (ProSe) with device-to-device communication for UEs in the neighborhood

Features

- VoWLAN (Voice over WLAN)
 - APIs for 3rd party access to 5G services
- Requirements for Mobile communication system for railways (Future Railway Mobile Communication System, FRMCS)
-

The NSA architecture (non-standalone) shown in Figure 6.1 characterizes the early drop. Here new NR base stations are used, i.e., the advantages of the new radio technology, such as the higher bit rates, can already be used. According to rapid 5G roll-out, these base stations, also referred to as en-gNB (next generation NodeB), will be operated on the existing 4G core network EPC (see Section 2.5) together with LTE-eNodeBs (eNB). The eNodeB is the master for signaling and serves as a mobility anchor; the UE uses dual connectivity (DC) to both NBs; the en-gNB acts as a booster.

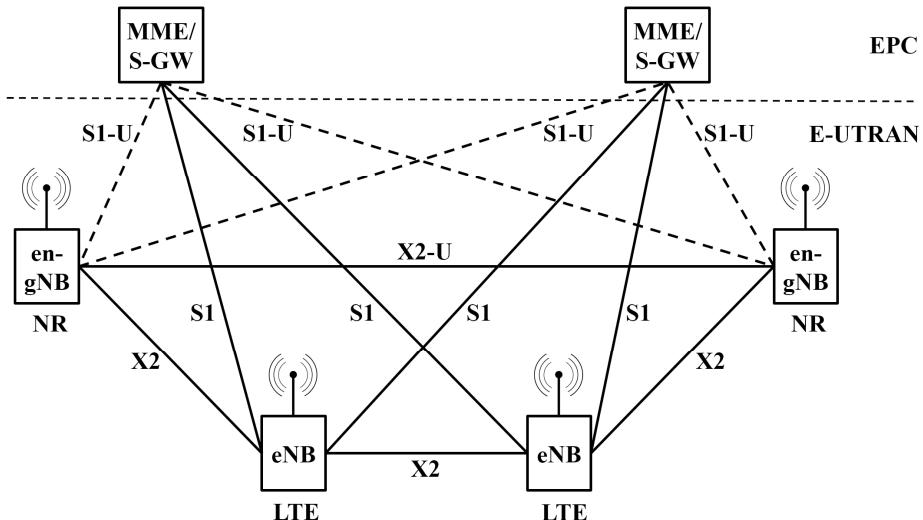
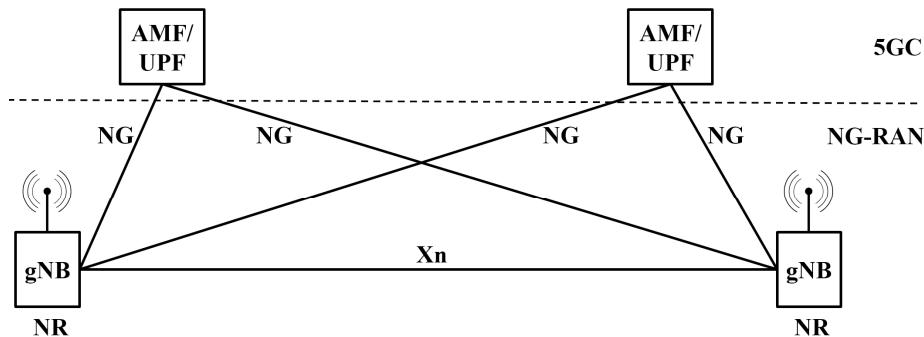


Fig. 6.1: 5G-NSA architecture [19]

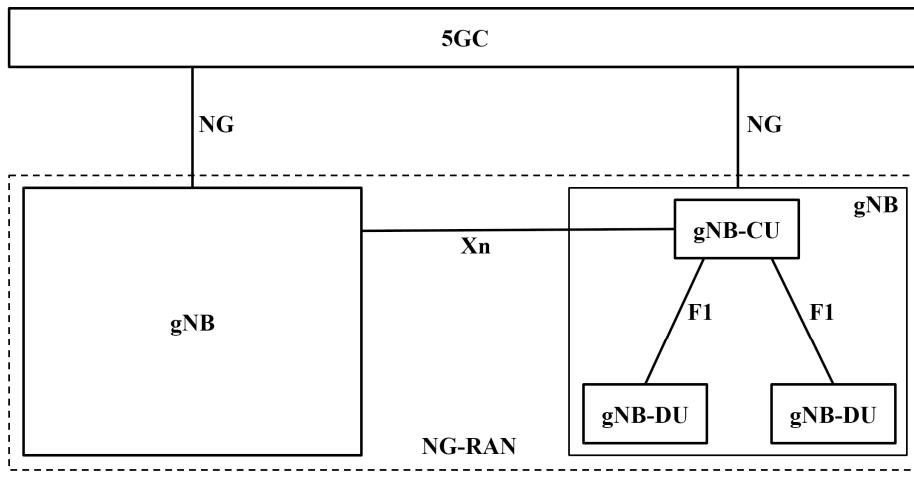
The main drop within Release 15 standardizes the SA architecture (standalone) and, thus, the use of a new core network, the 5G Core (5GC). As shown in Figure 6.2, the new 5G base stations gNB can now be connected directly to the 5GC, 5G operation is possible standalone, and no 4G infrastructure is necessary [19].



AMF = Access and Mobility Management Function UPF = User Plane Function

Fig. 6.2: 5G-SA architecture [19]

A 5G access network, the Next Generation RAN (NG-RAN), contains gNBs that are connected to the 5GC and possibly also to each other. According to flexibility in network design (radio technology on-site, remote control) and costs (central processing with C-RAN, see Section 3.1), a gNB is divided into a gNB control unit (CU) and one or more gNB distributed units (DU) as shown in Figure 6.3. In addition, the CU is divided into control and user plane functions because of modularization.



CU = Control Unit

DU = Distributed Unit

Fig. 6.3: NG-RAN architecture with split gNB [19]

According to Table 6.1 and Figure 6.4, the SBA is the basis of the 5G core network 5GC. In contrast to conventional monolithic network nodes, the necessary network

functions are provided by relatively fine-grained network functions (NF), which offer their services to other NFs via uniform interfaces within a framework. This ensures modularity, reusability, and flexible combination, an optimal basis for the use of NFV. As shown in Figure 6.4, the actual AF (Application Function) uses a combination of required NFs. This figure also shows the division of the network functions into a user plane for user data (UPF, User Plane Function) and a control plane for control and signaling (AMF, Access and Mobility Management Function; SMF, Session Management Function) [19].

SBA with NFV is an essential prerequisite for Network Slicing, i.e., the provision of various specialized logical networks based on physical network infrastructure. This allows virtual networks to be simultaneously implemented within a 5G network, e.g., for IoT with high connection density, for smartphones with high bit rates, and for V2X with short delay times and high availability.

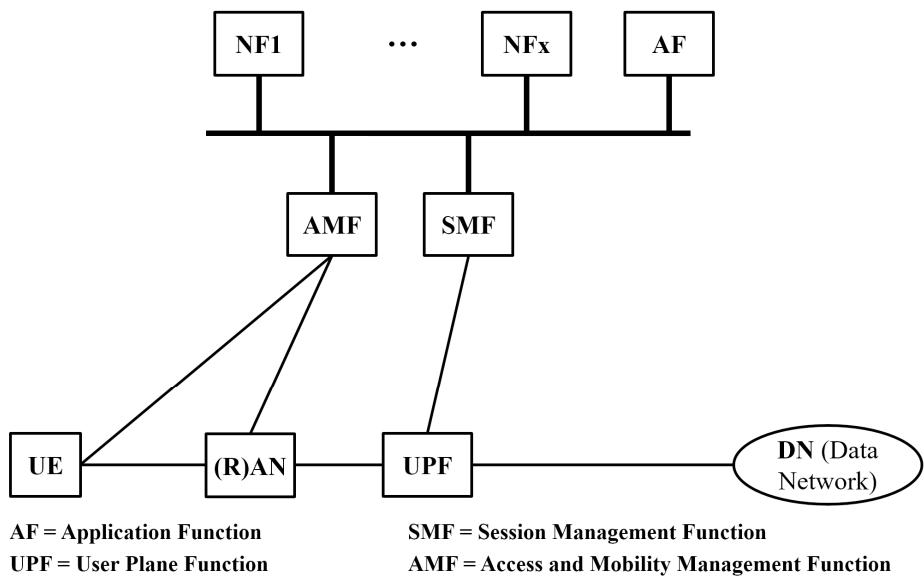


Fig. 6.4: 5G system architecture with SBA [19]

According to Table 6.1, Release 15 5GC also supports local hosting of services and edge computing based on MEC (see Section 3.1). This allows a service to be provided close to the user, e.g., on a base station, to achieve extremely short end-to-end delays in V2X.

The uniform access control provided by the 5GC can be used to decide whether to allow or deny UE access based on various combinable criteria such as operator

specifications, network expansion, user profile, or available services, e.g., in the event of congestion.

The 5GC supports any access network with 3GPP technology, i.e., 5G NR and 4G E-UTRA (LTE), and non-3GPP access, e.g., WLAN via the Internet, i.e., also from an insecure environment.

The 5GC also provides a framework for the session, access, and mobility control, as well as QoS and charging. The QoS support can be requested and supplied per flow and is not only applicable to the user data but also the signaling.

Furthermore, network functions can be made available to 3rd party providers, for example, to manage a customer-specific network slice (e.g., for Smart Grid) or an application hosted in the network (e.g., with frequently used high-resolution video streams).

Other 5G features provided by Release 15, such as vehicle-to-x communications, are listed in Table 6.1 above [19].

With its advanced functions, Release 16 standardizes a 3GPP 5G system phase 2 following Table 6.2, which in turn should cover the requirements of the ITU IMT 2020 target system. In the meantime, a successful evaluation has been carried out. The results of 3GPP are described in TR 37.910 [56; 62], submitted to the ITU, and incorporated in the higher-level ITU-R recommendation M.2150-1 [119]. This is subsequently referenced by the national regulatory authorities (see Section 5.3).

The most important innovations concern the NR radio interface with higher end-user bit rates by introducing different configurations of Carrier Aggregation (CA) and adding 256-QAM (Quadrature Amplitude Modulation) for FR2. Other enhancements include NR-based access to unlicensed spectrum and optimizations in mobility support and UE power consumption. The overall trend in Release 16 is to make it a communications platform suitable for various industries (verticals), such as transportation (autonomous driving, V2X, rail, shipping), automated factories, healthcare, public safety, and many more. In this regard, the versatility and reliability of the 5G system have been further enhanced to make it industry-compatible, with improvements for Ultra-Reliable Low Latency Communications (URLLC), network slicing, edge computing, Cellular IoT, Non-Public Networks, positioning and LAN-like services. In addition, the use of 5G as an underlying communications network (i.e., for transparent use by outside-the-network applications, 3rd party providers) has been improved, primarily as part of the work on northbound APIs. In addition to all these industrial aspects, other Release 16 enhancements address the coexistence of 5G with radio-based or wireline non-3GPP access network systems, entertainment (e.g., streaming and media distribution), and network optimizations (e.g., regarding user identity) [20; 60].

Tab. 6.2: Features and functionalities of 3GPP Release 16 [20; 60; 187]

Features
New
LAN support
TSN support (Time-Sensitive Networking) with highly accurate time synchronization
Vehicle-to-Everything (V2X)
Northbound APIs
ATSSS (Access Traffic Steering, Switch and Splitting)
Mobile Communication System for Railways (FRMCS) Phase 2
Support for campus networks (Non-Public Networks, NPN)
NR for high-speed trains
NR with unlicensed frequencies in 5- and 6-GHz bands
NR with FR2 frequency range of 39.5 – 43.5 GHz
NR-based radio interconnection of base stations (Integrated Access and Backhaul, IAB)
NR DL with 256-QAM transmission in FR2
Enhancements
URLLC support
NB-IoT
Coexistence with non-3GPP access network systems
Mission-critical communication
SRVCC (Single Radio Voice Call Continuity) from 5G to 3G (see Section 2.6)
Streaming
Location and position accuracy
Network slicing
SBA
IMS integration
Reduced power consumption for UEs
Dual Connectivity (DC) with lower activation times and higher data rates
Carrier Aggregation (CA)
SON (Self-Organising Networks) support for NR
NR functions, including those for MIMO antennas

Release 17 was finally standardized at the end of 2022. The new features and improvements are described in the associated TR 21.917 [243]. According to Table 6.3, the main innovations in Release 17 are new frequency ranges for the use of millimeter waves, 1024-QAM for DL transmission in NR FR1, communication via satellites, network slicing in the RAN, NR for UEs with lower complexity for energy and cost savings (reduced capacity (RedCap), e.g., AR glasses or generally wearables with

reduced bit rate, fewer receive antennas, and no dual connectivity support, etc.), and 8K TV support. In addition, Release 17 also includes numerous improvements and enhancements to previously available features [61; 188; 243].

Tab. 6.3: Features and functionalities of 3GPP Release 17 [61; 188; 243]

Features
New
NR with FR2 frequency range of 47,2 – 48,2 GHz and 57 – 71 GHz
NR DL transmission with 1024-QAM in FR1, e.g., for fixed wireless access
NR across Non Terrestrial Networks (NTN)
5G network components in satellites
IoT via NTN
RAN Slicing
NR für RedCap-UEs
8K TV support
Comprehensive teleconferencing and telepresence
Support for UEs with multi SIM cards
Enhancements
Dynamic spectrum sharing of a frequency range for NR or LTE according to the users' demand (Dynamic Spectrum Sharing (DSS))
NR carrier aggregation
ATSSS
NR for UEs with lower complexity (Reduced Capacity, RedCap), e.g., AR glasses with reduced bit rate and number of antennas
Network coverage
Device-to-Device and sidelink communication
IoT support
Handling of short data packets, e.g., of sensors, at mMTC (small data)
APIs for 3rd party users from different industries (verticals)
Support for drones, UASs, and UAVs
Non-Public Networks
FRMCS for railroads
V2X services
Network slicing
Edge computing
Multicast and broadcast traffic
Multimedia Priority Service (MPS) for prioritized communication of emergency organizations and security authorities, Mission Critical (MC) communication

Features

Self Organizing (SON)/autonomous network

Security

Network management

6.3 5G Network Architecture

Based on the requirements for a 5G system in Section 4.3, the desired features in Section 6.2, and in particular, the 5G design principles and key technologies in Section 6.1, the basic 5G network architecture is clearly and comprehensively described in Figure 6.5.

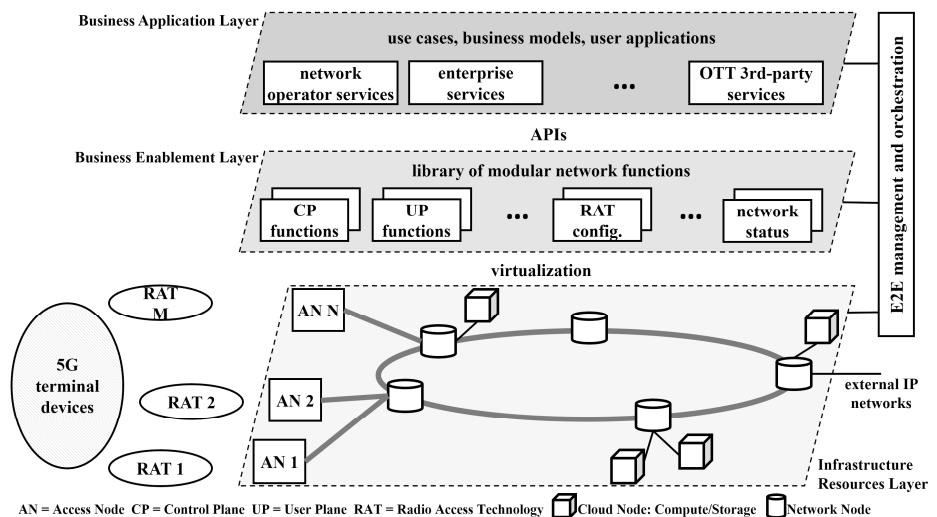


Fig. 6.5: 5G network architecture according to NGMN Alliance [96]

Due to the requirements, which in some cases vary greatly depending on the area of application, a 5G network must provide a wide variety of services at different locations with widely varying bit rates and numbers of connected terminals. This requires enormous flexibility, scalability, and elasticity. It can be best achieved with several application-specific virtual networks on a physical infrastructure based on NFV and SDN. The NGMN Alliance has taken up and elaborated this approach in [96]. The Infrastructure Resources Layer in Figure 6.5 provides the required physical network infrastructure with SDN-based switching network nodes and computing and storage power (e.g., in data centers). The various access networks, including

RATs, are connected to the core network infrastructure via access nodes (AN). The required network functions are taken from a library in the Business Enablement Layer and provided in the form of SW instances as virtual network elements on the cloud nodes in the Infrastructure Layer. Their functionality is then accessed by the 5G terminals, the RATs, and the network operator, enterprise, or even OTT-3rd party services (Over The Top) from the Business Application Layer. A comprehensive system for end-to-end management (E2E) and orchestration of hardware, software, and services ensures that end-to-end operations are automated and consistent across all three layers [96; 173].

Besides, the details shown in Figure 6.6 illustrate the advantages of such a 5G network architecture.

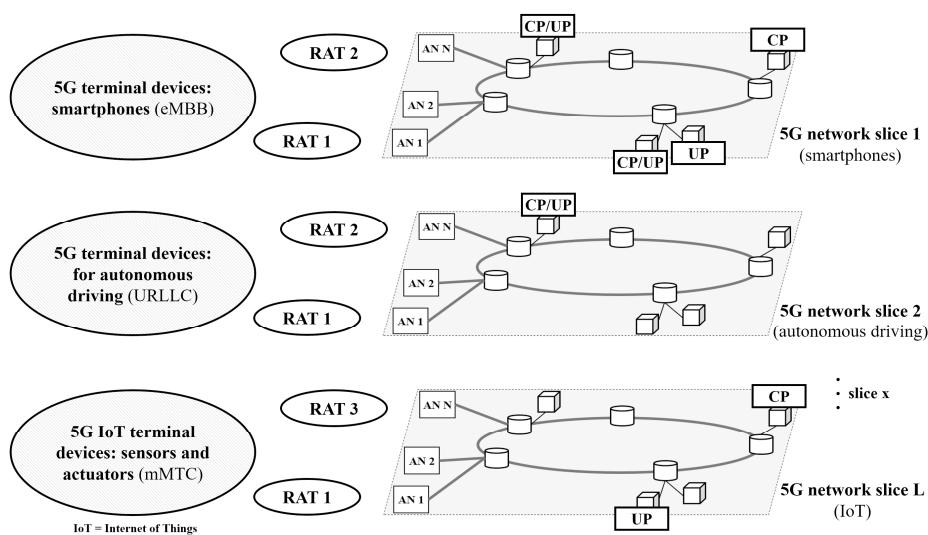


Fig. 6.6: Different 5G network slices based on the same physical infrastructure [96]

Here it becomes evident that by using NFV and SDN for different usage scenarios with various requirements such as smartphones, autonomous driving, or IoT, optimal virtual networks with the necessary network functions can be provided on a single physical platform. The individual so-called network slices (see Section 8.4) are extremely scalable, i.e., computing power, memory, virtual machines, and network functions can be switched on or off and/or moved as required [96; 173].

These two architectural views illustrate that such a 5G network implements the design principles mentioned above: All-IP network, modularization with SBA, network softwarization with NFV, SDN and network slicing, multi-tenant capability, cloudification including C-RAN and MEC, openness for 3rd party providers, heterogeneous RAN technology, various radio and wired access network technologies,

core network decoupled from access network technology, flexible terminal device connectivity, downward and upward compatibility.

7 5G Access Networks

The great challenges to the 5G access networks were already apparent in Section 4.3 for the 5G requirements with the comparatively high bit rates and Section 5.1 for the possible, even completely new frequency spectra. Therefore, this topic shall be discussed further here by dealing with the radio transmission technology in Section 7.1 and in Section 7.2 with possible RAN architectures and the corresponding functions. Section 7.3 presents an introduction to Open-RAN.

7.1 Radio Transmission Technology

LTE had and has its focus on Mobile Broadband use cases (MBB), extended by MTC and narrowband IoT. 5G extends the two application areas mentioned above to higher bit rates (eMBB) and higher connection densities (mMTC, whereby NB-IoT and LTE-M is used). It introduces URLLC with very low latency and very high availability for function-critical applications (see Chapter 4).

The radio transmission technology for NR is based on LTE concepts. But it optimizes and expands these concepts due to the increased requirements for performance indicators and flexible handling, e.g., in the various frequency ranges used (see Section 5.1).

For a more concrete understanding, we will give an overview of the functions, interrelationships, and operating modes of the physical layer at NR. For data transport via the NR radio interface, the following functionalities are required:

- Error detection on the transport channel and indication to higher layers
- FEC encoding/decoding (Forward Error Correction) of the transport channel
- Hybrid ARQ procedure (Automatic Repeat Request) for resending lost messages
- Rate matching of the coded transport channel to physical channels
- Mapping of the coded transport channel onto physical channels
- Power weighting
- Modulation and demodulation
- Frequency and time synchronization
- Radio characteristics measurements and indication to higher layers
- MIMO (Multiple Input Multiple Output) antenna processing
- RF processing (Radio Frequency) [49].

Compared to LTE, the NR radio interface presents particular challenges due to the very different frequency ranges to be supported from 600 MHz to 28 GHz and more (see Tables 5.1 and 5.2), the possibly very high bit rates up to 20 Gbit/s, the very low latency down to 1 ms as well as very low power consumption in IoT applications and

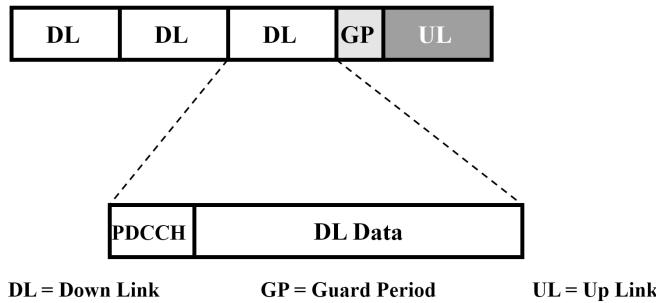
due to energy efficiency. The specification of the NR radio interface had to take all these into account.

Three physical channels each are transmitted in the downlink (DL, gNB → UE) and uplink direction (UL, UE → gNB). In DL, these are

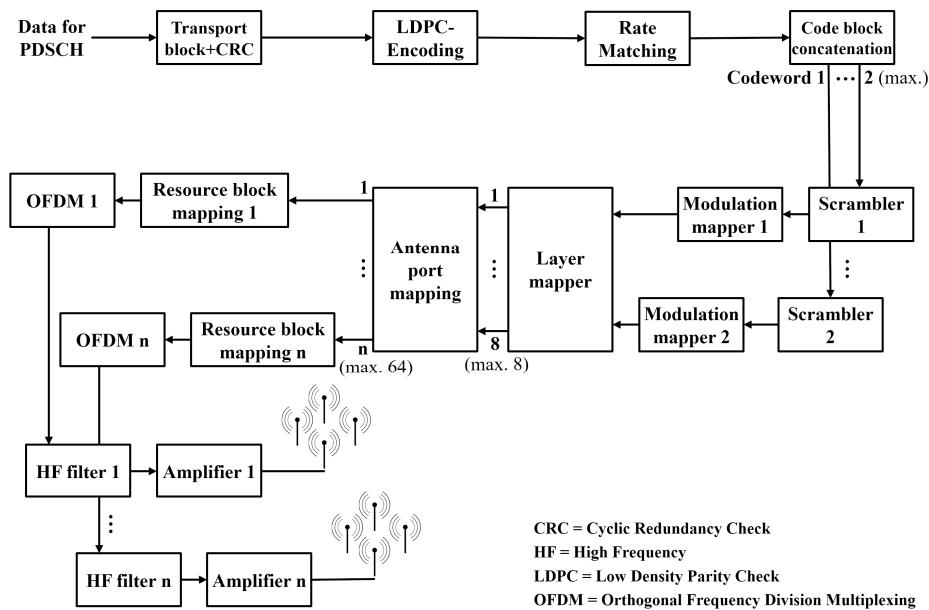
- Physical Downlink Shared Channel (PDSCH),
- Physical Downlink Control Channel (PDCCH),
- Physical Broadcast Channel (PBCH),
in UL
- Physical Random Access Channel (PRACH),
- Physical Uplink Shared Channel (PUSCH) and
- Physical Uplink Control Channel (PUCCH) [48].

The PDCCH transmits control information from the base station to the terminals and, among others, allocates the necessary resources for the PDSCH and PUSCH. The PDSCH represents the actual transmission channel for data transfer from the gNB to the UE. Paging, the calling of a UE in the radio cell, also takes place via this channel. Finally, the PBCH provides a periodically transmitted broadcast signal, supporting the UEs' access to the NG-RAN. In the UL, the PUCCH is responsible for the transmission of the control information, including the HARQ feedbacks (Hybrid Automatic Repeat Request). The data transfer from the UE to the gNB takes place in the PUSCH, while the PRACH enables the procedure for the random access of the UEs to the NG-RAN, e.g., in case of a handover. Also, the DL and UL transmit further required reference and synchronization signals [106].

According to Table 5.1, the transmission between gNB and UE uses either Frequency Division Duplex (FDD) or Time Division Duplex (TDD) to separate DL and UL. FDD transmits in DL and UL direction in two different frequency ranges. TDD transmits in only one frequency range alternately at other times. FDD is advantageous for VoIP or MoIP with equal bit rates in DL and UL because of the same bandwidth in both directions. TDD is useful for services with asymmetric bit rates, such as web browsing because of the possible flexible distribution of transmission resources between DL and UL. Figure 7.1 gives an example of TDD with DL and UL phases, where a guard period must be kept before switching to the UL phase. A disadvantage of TDD is a possible interference, the so-called cross-link interference, from a neighboring base station transmitting during a receive phase that is not synchronized in time [91].

**Fig. 7.1:** Example of a TDD signal structure

For the provision of the different physical channels mentioned above, the same functions are required in principle. However, there are differences according to the tasks. As an example, we describe the Physical Downlink Shared Channel, the PDSCH for the transport of data from the gNB to the UE, which is essential for the communication services, in more detail using the block diagram in Figure 7.2.

**Fig. 7.2:** PDSCH transmitter [106]

The data to be transmitted in the PDSCH, e.g., to a UE, are in a first step segmented into transport blocks and cyclically supplemented with test data using the CRC pro-

cedure (Cyclic Redundancy Check). This enables the detection of errors on the receiver side. Subsequently, redundant information is added to the data blocks through channel coding to not only detect errors in the receiver but also to correct them. The PDSCH uses the FEC block code (Forward Error Correction) LDPC (Low Density Parity Check) to ensure high bit rates due to its lower complexity (compared to the Turbo Code used in LTE) and its high performance. The LDPC-coded data is then adapted to the bit rate available in the radio channel according to an algorithm, also known in the receiver (rate matching). Bits suppressed by the rate matching algorithm on the transmitter side are added back in the receiver.

For the transmission of the transport blocks, one or two parallel transmitted data sequences can be used per UE in the DL (in the UL, only one). These are designated with codeword 1 (CW) and 2. In the variant with 2 CWs, each transport or code block from the concatenated sequence is divided into 2 blocks, which can then be coded and modulated separately. This increases flexibility in terms of adapting to the radio channel. Besides, the bit rate is halved by splitting a transport block into 2 CWs. That means, with 2 CWs, we will achieve double data throughput.

Further following Figure 7.2, the coded transport blocks are scrambled to ensure sufficiently frequent changes in the data sequence for the demodulation. In the next step, the bits to be transmitted are combined into symbols according to the modulation method used. In the case of QPSK modulation (Quadrature Phase Shift Keying), due to long-range with high attenuation and/or low quality of the radio channel, 2 bits are mapped to a symbol, here a phase shift (Modulation Mapper): $00 \rightarrow 45^\circ$, $01 \rightarrow 315^\circ$, $10 \rightarrow 135^\circ$, $11 \rightarrow 225^\circ$. In the actual implementation, the time signal – as shown in Figure 7.3 – is represented by corresponding complex symbols with in-phase and quadrature components. In 256-QAM (Quadrature Amplitude Modulation) for very high bit rates, 8 bits form a symbol, where such a symbol represents one of 256 amplitude-phase combinations [106; 91].

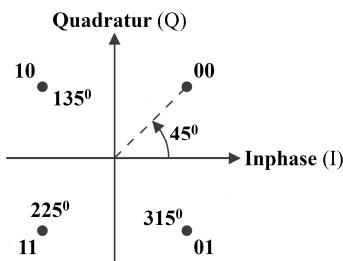


Fig. 7.3: QPSK [91]

Subsequently, the up to 2 symbol sequences (1 or 2 CW) are mapped to up to 8 streams or layers (Layer Mapper). For only one CW, 1 to 4 layers are used, with 2

CWs at least 5 to 8 layers. The number of layers used indicates the achievable degrees of freedom for MIMO transmission (Multiple Input Multiple Output) with several antenna systems, i.e., how many independent radio channels can be realized. In the case of transmission for interference reduction, for example, with 2 transmitting antennas (diversity), the symbols of a codeword can be mapped individually to 2 layers, which in turn are mapped to the antenna ports after a precoding process. If there are more layers, the symbols are nested across the individual layers. With NR, there are up to 8 layers per UE in DL and 4 in UL (Single User MIMO) and up to 12 layers in DL and UL if several UEs are served (Multi User MIMO), e.g., 6 UEs with 2 layers each at one gNB.

The up to 8 layers are again mapped to antenna connectors using precoding (antenna port mapping). This step is necessary because there are usually more antenna connectors than layers or MIMO systems: e.g., 8 layers with 8 antenna connectors each, i.e., 64 ports and even more antenna elements, e.g., 192. The reason for the latter is that by combining several antenna elements (e.g., dipoles) per port, we can achieve the desired beam pattern.

For each antenna port, the necessary resource blocks are assigned according to the required bandwidth (resource block mapping). This is followed by the modulation and formation of the OFDM signal (Orthogonal Frequency Division Multiplexing) for each port, as shown in Figure 7.2. The generated analog RF signal (Radio Frequency) has to be filtered and amplified before transmitted via the corresponding antenna elements [106; 91].

The types of modulation used in OFDM are usually in DL and UL QPSK, 16-QAM, 64-QAM, or 256-QAM. The more unproblematic the radio channels and the higher the desired bit rates are, the more transmit symbols are used, from in the worst case 4 with QPSK up to 256 with 256-QAM [48]. As of Release 17, 1024-QAM can also be used in the downlink [243].

The modulation process for an NR transmitter is explained below using the comparatively simple quaternary QPSK procedure. According to Figure 7.3, 2 bits each are mapped to a symbol, here a phase shift: $00 \rightarrow 45^\circ$, $01 \rightarrow 315^\circ$, $10 \rightarrow 135^\circ$, $11 \rightarrow 225^\circ$. In practice, the symbol sequence is modulated digitally, i.e., using a signal processor, with an intermediate frequency (f_{if}) according to the simplified representation in Figure 7.4. The result is a sine signal $s_{\text{if}}(t) = S \cos(2\pi f_{\text{if}} t + \phi(t))$ according to the symbol sequence changing phase ϕ . This digitally modulated signal is converted into an analog signal through a D/A converter (Digital/Analog) and, in a second step, modulated with the high-frequency carrier for the radio channel. The result is a high-frequency sine signal $s_{\text{hf}}(t) = S \cos(2\pi(f_{\text{if}} + f_{\text{hf}})t + \phi(t))$ with the desired phase changes representing the bit combinations to be sent. As shown below, this two-step approach is very advantageous for the realization of OFDM [91].

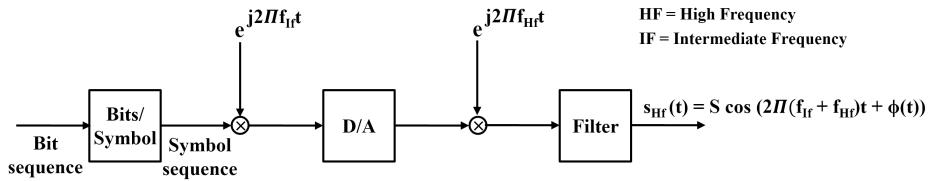


Fig. 7.4: QPSK modulation

Since a base station communicates with several terminals, a multiple access method is required for this 1-to-N situation. As with LTE or WLAN, the OFDMA (Orthogonal Frequency Division Multiple Access) method is used for this purpose. It uses OFDM (Orthogonal Frequency Division Multiplexing), i.e., frequency multiplexing using several or many orthogonal carrier signals (OFDM subcarriers) when modulating a data signal to be transmitted. Orthogonal means that a carrier signal has its maximum amplitude when the neighboring carriers have zero points. This minimizes the disturbing interference. If the OFDM subcarriers are applied not only to one but to several data signals, this is called OFDMA.

OFDMA or OFDM thus supports time-parallel communication between a gNB and several UEs, provides high spectral efficiency (in bit/s/Hz), and, in conjunction with channel coding, offers mechanisms to minimize fading (degradation of the received signal due to interference, shadowing, multipath propagation, and Doppler effect) and intersymbol interference due to multipath propagation and mobility [91].

Figure 7.5 explains the basic principle of OFDM. An OFDM transmitter takes a serial symbol sequence, performs a serial/parallel conversion, and transmits each of these individual symbols parallel on a different carrier frequency. These carrier frequencies are close together and are called sub-carriers. The required bandwidth and bitrate per subcarrier are low, but the symbol rate is still the same. One possibility shown in Figure 7.5 is a frequency spacing of 15 kHz between the subcarriers, which results in a symbol spacing according to $T_s = 1/\Delta f$ of 66,7 µs because of the desired orthogonality of the carriers.

An advantage of OFDM is not only that the requirements on the transmission channel per subcarrier are significantly reduced due to the comparatively low symbol rate. Additionally, particularly disturbed frequency ranges with still free subcarriers can simply be omitted, i.e., the associated subcarriers are not used.

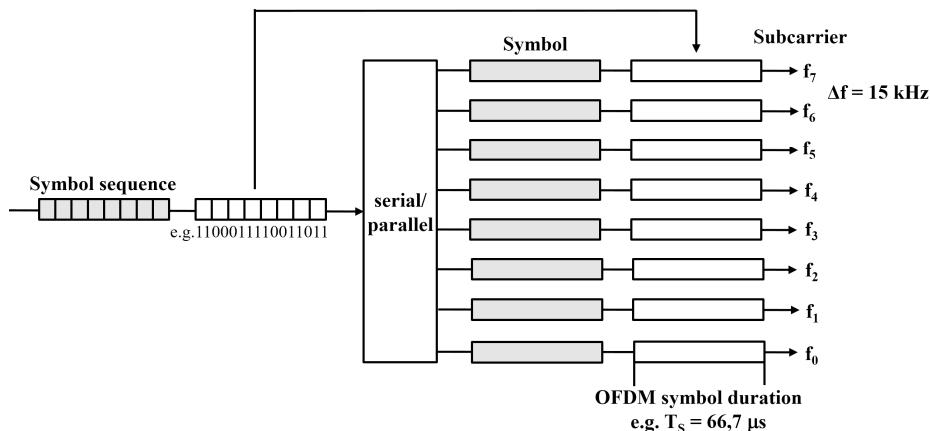


Fig. 7.5: OFDM basic principle [91]

The operation of an OFDM transmitter is concretized by Figure 7.6. A sequence of 8 QPSK symbols is converted serially/parallel, and QPSK modulated with 8 intermediate frequencies at 15 kHz intervals. The individual modulated signals are added up and normalized. The sum signal, the resulting OFDM symbol, is D/A converted and then modulated for the radio channel with the high-frequency carrier signal, then filtered, amplified, and transmitted. As already mentioned, depending on the requirements and the radio channel, 16-QAM, 64-QAM, or 256-QAM can be used in addition to QPSK, and from Release 17, also 1024-QAM in the downlink.

The process of parallel modulation with closely adjacent intermediate frequency subcarriers can be described mathematically by an IFFT (Inverse Fast Fourier Transformation) and is, therefore, comparatively inexpensive and straightforward to implement with a signal processor. For OFDM decoding, an FFT (Fast Fourier Transformation) must be applied in the receiver [91; 106].

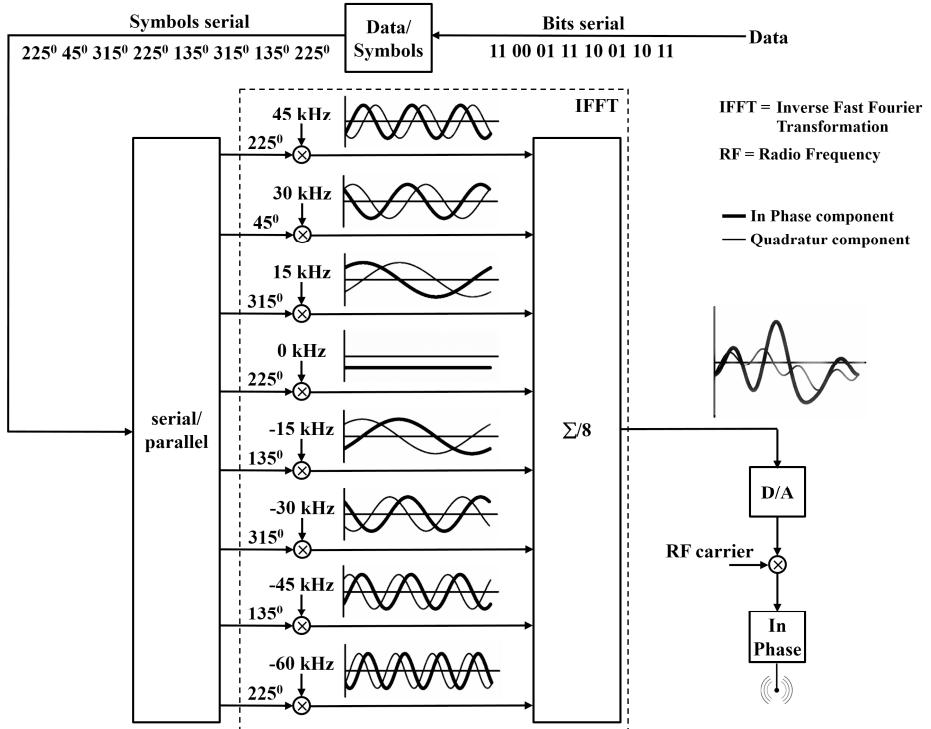


Fig. 7.6: OFDM processing [91]

The OFDM symbols obtained by summing the subcarrier signals transport not only the PDSCH for the actual data traffic but also the other physical channels, such as the PDCCH control channel, and the reference and synchronization signals, such as the DM-RS (Demodulation-References Signal). For demultiplexing these channels in the receiver, a frame structure is required for the transmission. As shown in Figure 7.7, it normally consists of 14 OFDM symbols, forming a slot (time slot). Depending on the subcarrier frequency spacing, 1 (15 kHz), 2 (30 kHz), 4 (60 kHz), 8 (120 kHz), or 16 slots (240 kHz) form a 1 ms subframe. Besides, a CP (Cyclic Prefix), a copy of the last part of the OFDM symbol, is added to each OFDM symbol to ensure the orthogonality of the subcarriers. Thus, it enables the receiver to distinguish between successive OFDM symbols despite multipath propagation with resulting interference. Therefore one speaks of CP-OFDM.

It should also be mentioned that NR does not only use classical CP-OFDM but Filtered-OFDM (F-OFDM). Additional filters can be placed over specific subcarrier blocks to adapt to different usage scenarios.

The duration of a CP is selected depending on the subcarrier frequency spacing. For example, for 15 kHz, a symbol spacing of $T_s = 1/\Delta f$ of 66,7 µs results in a CP dura-

tion of 4.7 μ s, a total of approx. 71.4 μ s per symbol and 1 ms for 14 symbols per slot. For larger subcarrier frequency distances, the values are correspondingly smaller. To achieve short delay times even at 15 kHz, e.g., for IoT applications, mini-slots with only 2, 4, or 7 symbols can be formed. 10 subframes together form a 10 ms frame [106].

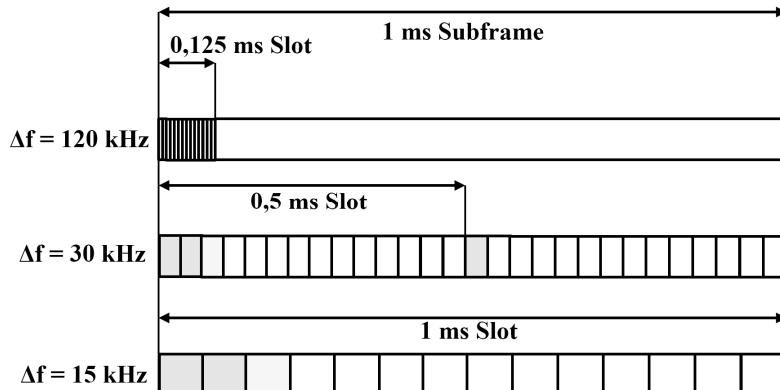


Fig. 7.7: Frame with OFDM symbols [106]

The transmission of the modulated signal, which is a sequence of OFDM symbols, requires resources. If this is a single subcarrier, it is called a Resource Element. 12 consecutive subcarriers for transferring one slot, i.e., 14 successive OFDM symbols, are combined into one Resource Block (RB). Thus, the number of RBs and the subcarrier frequency spacing give the available transmission bandwidth, or vice versa, for a required bandwidth, one can determine how many RBs of which subcarrier type are needed. Table 7.1 illustrates the relationships based on TR 38.211 [50], whereby the contents apply equally to the DL and UL directions. However, in newer versions of TR 38.211, the number of RBs is handled more flexibly.

Using Table 7.1, you can obtain the bandwidth when using 24 RBs with 15 kHz subcarriers from $24 \cdot 12 \cdot 15 \text{ kHz} = 4,32 \text{ MHz}$. With 275 RBs and 60 kHz subcarriers, the result is $275 \cdot 12 \cdot 60 \text{ kHz} = 198 \text{ MHz}$. Conversely, for a bandwidth of approx. 20 MHz, using 30 kHz subcarriers, about 56 RBs are necessary according to $20 \text{ MHz}/(12 \cdot 30 \text{ kHz})$.

Tab. 7.1: Subcarrier frequency spacing, resource blocks, and bandwidth at NR [50]

Δf [kHz]	OFDM symbols/ slot	Slot/ sub-frame	Slot/ frame	Minimum number of RBs	Maximum number of RBs	Minimum bandwidth [MHz]	Maximum bandwidth [MHz]
15	14	1	10	24	275	4,32	49,5
30	14	2	20	24	275	8,64	99
60	14	4	40	24	275	17,28	198
120	14	8	80	24	275	34,56	396
240	14	16	160	24	138	69,12	397

Below 6 GHz (FR1), subcarrier frequency spacings of up to 60 kHz are used. This allows bandwidths up to 200 MHz, although currently, only a maximum of 100 MHz is used. Above 24 GHz (FR2), subcarriers with frequency spacings of 60 kHz and above can be used, allowing bandwidths up to 400 MHz [106].

The number of bits transferred per subcarrier depends on the modulation method used. Table 7.2 shows the relationships. Considering the information in Tables 7.1 and 7.2, one can roughly calculate the gross bit rates we can achieve. With a bandwidth of 100 MHz, 30 kHz subcarrier frequency spacing, and 256 QAM, 275 slots with 14 OFDM symbols, each with 12 subcarriers for 8 bits each, are transmitted in a period of 0.5 ms according to the 275 RBs then used. Accordingly, a peak bit rate of $(275 \cdot 14 \cdot 12 \cdot 8)$ bit/0,5 ms = 0,74 Gbit/s is reached in this case. If not only 1 but 4 such streams are transmitted simultaneously using MIMO technology (4 x 4 MIMO), the bit rate quadruples to $4 \cdot 0,74$ Gbit/s = 2,96 Gbit/s. [131] describes the procedure for a more precise calculation of the throughput with NR.

Tab. 7.2: Modulation method and number of bits transmitted

Modulation method				
	QPSK	16-QAM	64-QAM	256-QAM
bit/modulation symbol	2	4	6	8
bit/RB	24	48	72	96

The radio transmission technology used by NR enables a very flexible usage according to the supported services with different bit rates, latency, frequencies, cell sizes, and radio channel qualities.

As already mentioned, the use of massive MIMO technology also contributes to this. It means that several antenna systems are used for transmitting and/or receiving the radio signals. A significant advantage has already been mentioned, the mul-

tiplex gain and, thus, higher data rates through MIMO. As discussed above and shown in Figure 7.2, a data sequence to be transmitted can be divided into several streams, each of which is then transmitted via a separate antenna or antenna system. In the case of N antennas (e.g., 8), the total bit rate is then N times as high as the data rate of the individual streams with the advantage of lower demands on the radio transmission technology due to the lower data rate by approx. $1/N$. This is one of the reasons for the very high bit rates possible with NR. This technique, called Spatial Multiplexing or MIMO-Space Multiplexing, with several data streams transmitted via MIMO, can be used between two individual systems, e.g., between a UE and a gNB. This is called Single User-MIMO. If the N streams (e.g., 2) come from different transmitters or are sent to different receivers, it is a Multi User-MIMO.

Also, MIMO's application, the transmission, and/or reception with several antennas provide a so-called group gain, a higher receive power. More receiving antennas, in sum, take up a higher receiving power than a single antenna due to the superposition. Accordingly, the radio channel attenuation can be higher. Besides, MIMO in the receiver can be used to reduce intersymbol interference caused by reflections and shadowing due to the multipath propagation of radio signals. Last but not least, the use of MIMO antenna systems brings a diversity gain. The radio signal cancellation due to delayed, reflected, and then superimposed signals can be counteracted.

Finally, the MIMO technology is used for beamforming, which is crucial for NR at higher frequencies, already from 3.4 GHz, because of the higher attenuation. Appropriate control of the antennas ensures a targeted alignment of the antenna radiation or a constructive intersymbol interference by phase shifting, i.e., an advantageous superimposition of the radio waves at the receiver. The transmitting power is focused, the antenna gain is higher, a higher attenuation can be bridged. Figure 7.8 shows an example of the radio transmission from one gNB to two UEs [91]. If the individual beams in a radio cell point in different directions, the great advantage is that the same frequency resources can be used several times.

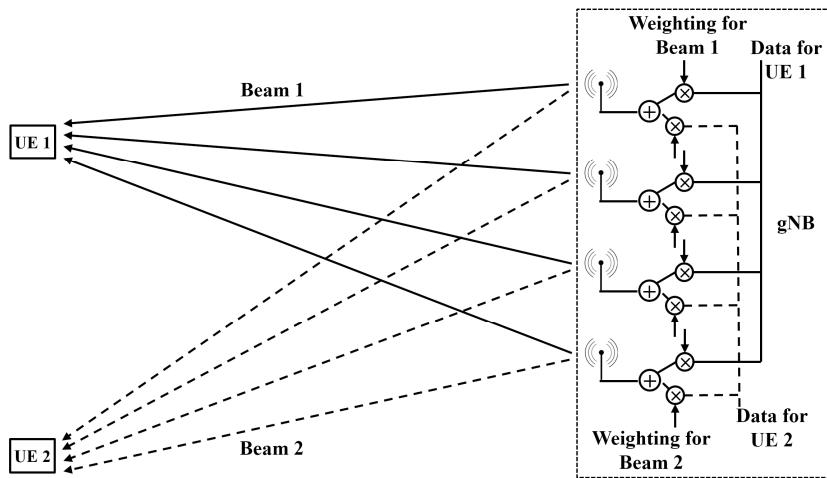


Fig. 7.8: Beamforming [91]

For beamforming, an alignment with feedback from the receiver must take place. Therefore, special subframes are sent. Based on the feedback, the most suitable beam can be determined. This beam sweeping process takes place in both DL and UL direction, as shown in Figure 7.9 [106].

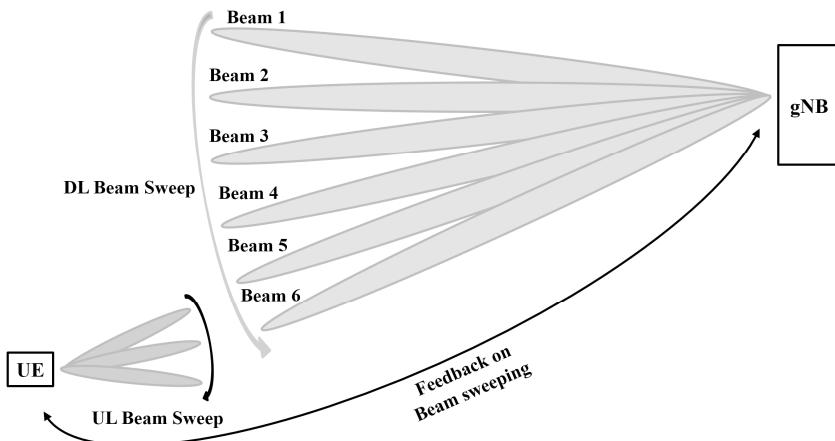


Fig. 7.9: Beam Sweeping [106]

7.2 RAN (Radio Access Network)

As already mentioned in Section 6.2 concerning the three successive Release 15 versions, a 5G-RAN distinguishes between the NSA (see Figure 6.1) and the SA solutions (see Figure 6.2). Besides, TR 23.799 [40] and others have worked out further RAN options, shown in Figure 7.10. Here you will find two standalone options and three non-standalone options. Section 6.2. has already discussed Option 3 with a conventional 4G core EPC (see Section 2.5) and connection of the 5G RAT NR via a 4G access network E-UTRA and Option 2 with new 5G NGC (Next Generation Core, 5GC) and new 5G access technology NR.

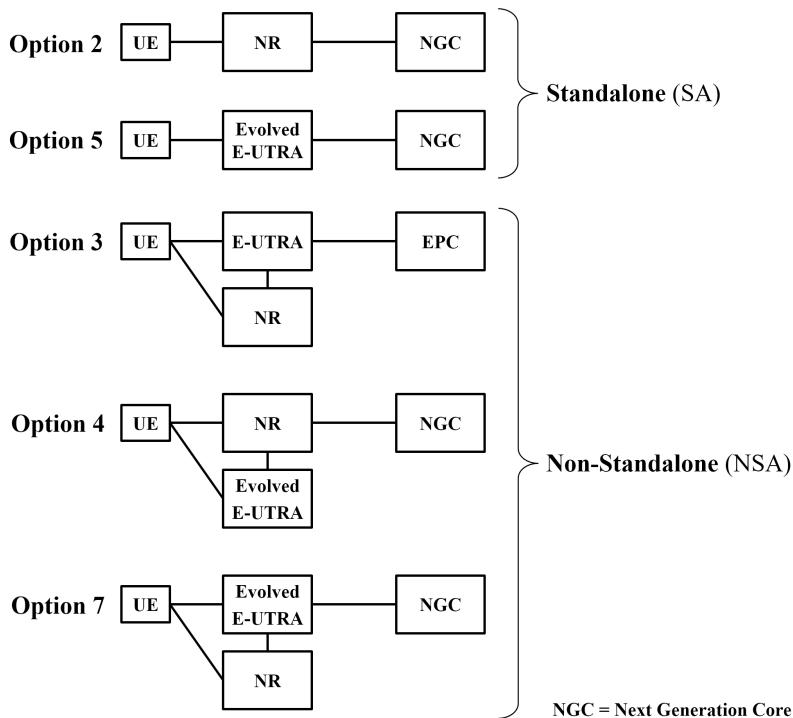


Fig. 7.10: RAN options [40]

The additional Option 5 in Figure 7.10 offers the possibility to operate a conventional 4G access network E-UTRA, called Evolved E-UTRA, after the upgrade, on a new NGC.

For the non-standalone architectures (in addition to the already mentioned Option 3 for fast entry of an incumbent operator into 5G), there are Options 4 and 7 for parallel operation of 5G and 4G RAT, starting from Option 2 or Option 5.

When introducing 5G, network operators have a choice of different options. They will choose them according to their initial situation, schedule, and expansion plans, including the planned user services. A start with Option 3 with an existing 4G network leads to a short-term 5G presence on the market and requires only comparatively low investments. However, only higher bit rates and thus only eMBB services are supported, and the necessary investments for further expansion and migration to an NGC are comparatively higher. A start with Option 2 directly offers the 5G service mix from eMBB via URLLC to mMTC but requires much higher investments at the beginning. However, the migration to a pure 5G system, which is necessary with a 4G network operated in parallel, is associated with comparatively much lower costs [40; 169].

In Figure 7.10, we use the terms from [40] to describe the 5G-RAN architecture options. In contrast, Figures 6.1 to 6.3 are based on the terms for the network elements and systems from [19]. For a better understanding of the following considerations and the content of the corresponding 3GPP-5G specifications and reports, Table 7.3 compares the different terms used in 3GPP. The gNB (next generation NodeB) is a 5G base station, also known as NR (New Radio), with a direct interface to the 5G core (5GC), also known as NGC (Next Generation Core). An en-gNB (E-UTRA-NR-gNB), on the other hand, represents a 5G base station with an interface to the 4G core (EPC) via a 4G base station eNB (evolved NodeB) with LTE functionality. An eNB is an LTE base station with an EPC interface, while an ng-eNB (next generation eNB) is an LTE base station with an interface to 5GC. In summary, the base station variants gNB (5G) and ng-eNB (4G) are connected to a 5GC, while eNB (4G) and en-gNB (5G) are connected to a 4G core network EPC [129; 40].

Tab. 7.3: RAN network options with 5G and 4G base stations [129; 40]

Option	Core network	Master Node (Base station)	Master RAT	Secondary Node	Secondary RAT
2	NGC = 5GC	gNB (next generation NodeB)	NR	-	-
5	NGC = 5GC	ng-eNB (next generation eNB)	Evolved E-UTRA	-	-
3	EPC	eNB (evolved NodeB)	E-UTRA	en-gNB (E-UTRA-NR-gNB)	NR
4	NGC = 5GC	gNB	NR	ng-eNB	Evolved E-UTRA
7	NGC = 5GC	ng-eNB	Evolved E-UTRA	gNB	NR

Figure 7.10 and Table 7.3 also show that there are 5G-RAN architectures, Options 3, 4, and 7, which combine two radio access technologies (RAT), 5G NR and LTE. The

changes introduced in LTE base stations for options 4 and 7 are 5G-specific. In other words, despite LTE radio technology, these are 5G base stations.

Independent of this, in all three cases, there is a Master Node, which is directly connected to the corresponding core network, and a Secondary Node (Slave), which is only indirectly connected via the Master Node. These constellations support Multi-Radio Dual Connectivity (MR-DC), i.e., a UE can not only communicate via one of the two RATs but also simultaneously via both. If this is done at the EPC with eNB as master and en-gNB as a slave, it is called E-UTRA-NR Dual Connectivity (EN-DC). If the two RATs are operated on a 5GC, three dual connectivity cases can be distinguished:

- NG-RAN E-UTRA-NR Dual Connectivity (NGEN-DC) with ng-eNB as master and gNB as a slave
- NR-E-UTRA Dual Connectivity (NE-DC) with gNB as master and ng-eNB as a slave
- NR-NR Dual Connectivity (NR-DC) with gNB as a Master and gNB as a slave.

In all four cases of MR-DC, a UE can use the resources of both RAT accesses [46].

If we sum up the mentioned RAN interconnection variants and consider the modularization options for a 5G base station (gNB) described in [19] and Section 6.2, a RAN architecture, as shown in Figure 7.11, is obtained. This takes into account the fact that a gNB can be arranged as a central gNB-CU (Control Unit) with distributed gNB-DUS (Distributed Unit) and that, unlike the 4G-RAT, a CU can be divided into CU-UP (User Plane) for user data handling and CU-CP (Control Plane) for signaling and control. Besides, it is possible to operate the Radio Unit (RU), which contains the radio transceiver with transmitting and receive amplifier, remotely at an antenna location far away from the base station. This 5G-RAN approach with a modular structure provides a high degree of flexibility and creates enormous possibilities for the optimal design of a 5G access network from both a technical and an economic point of view [19; 129].

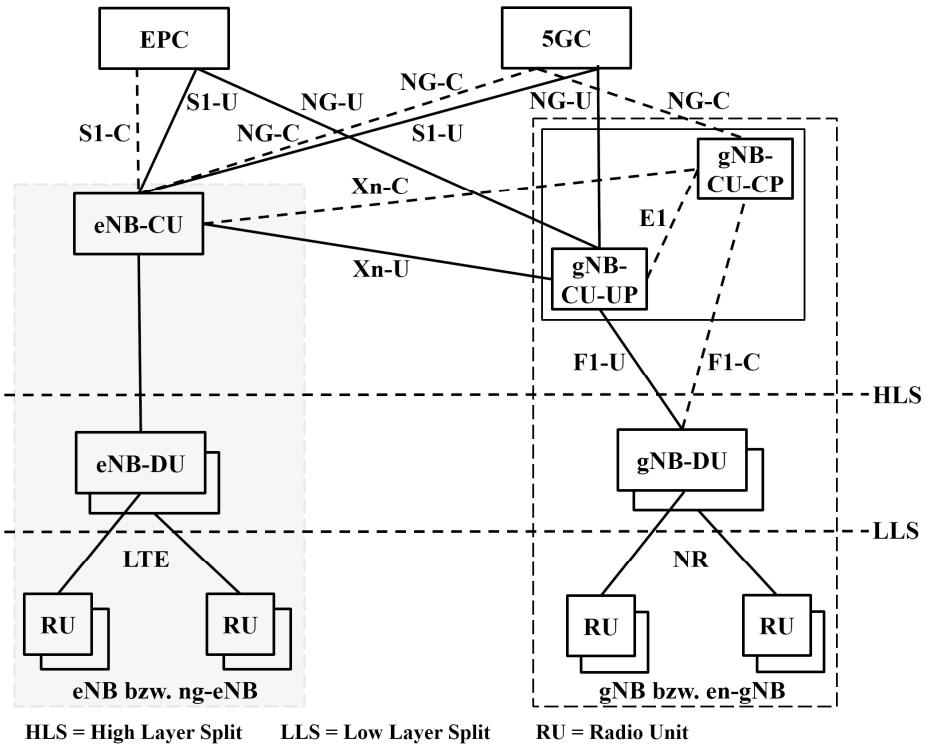


Fig. 7.11: RAN architecture for 5G, 4G and 5G/4G [129; 19]

In addition to the antennas, the Radio Unit (RU) comprises the specific radio interface hardware for modulation, digital/analog conversion, filtering, and signal amplification. A Distributed Unit (DU), which can serve several RUs connected via optical fibers in a so-called system area, contains the L1 baseband functions, the MAC protocol (Medium Access Control), and the RLC (Radio Link Control) handling, including MIMO (Multiple Input Multiple Output) and beamforming control. The CU-UP terminates the PDCP (Packet Data Convergence Protocol), provides message encryption, and controls dual connectivity. Its functionalities can be virtualized in terms of performance requirements. The CU-CP represents the RCF (Radio Control Function) for load sharing between system areas and different RATs, QoS negotiation, and overall RAN performance management. These functions are also very well suited for virtualization. In this respect, a more or less centrally located gNB-CU can serve many distributed and remotely operated gNB-DUs. It results in a cost-optimized solution and, thanks to the virtualization options, also offers the advantage of implementing gNB CUs in the cloud and thus cost-effectively as a C-RAN (see Section 3.1) [175].

Based on these considerations, Figure 7.12 shows different possibilities for splitting RAN functions and their spatial arrangement. A location can be at the antennas supplying the actual radio cell, a building of the network operator in the area of the access network (aggregation), or at the transition to the core network (edge). The remote operation of the RUs is called Low Layer Split (LLS), and that of DU and/or CU is called High Layer Split (HLS). According to Figure 7.12, the possible RAN architectures range from a remote RU, a remote DU operation, a distribution of CU, DUs, and RUs to three sites, additional remote CUs up to a monolithic, complete base station at the antenna site. Additionally, Figure 7.12 illustrates the possibility of edge computing to host a user application (app) in the RAN (see MEC in section 3.1) [129].

In summary, it is evident that the 5G design principles from Section 6.1 are also applied in the RAN as far as possible: All-IP network, modularization (RU, DU, CU-UP/CP), network softwarization (NFV), cloudification (C-RAN), openness for 3rd party providers (MEC), heterogeneous RAN technology (NR and LTE, etc.), various radio and wire-based access network technologies, core network decoupled from access network technology, flexible terminal device connectivity (dual connectivity) as well as downward (LTE, etc.) and upward compatibility.

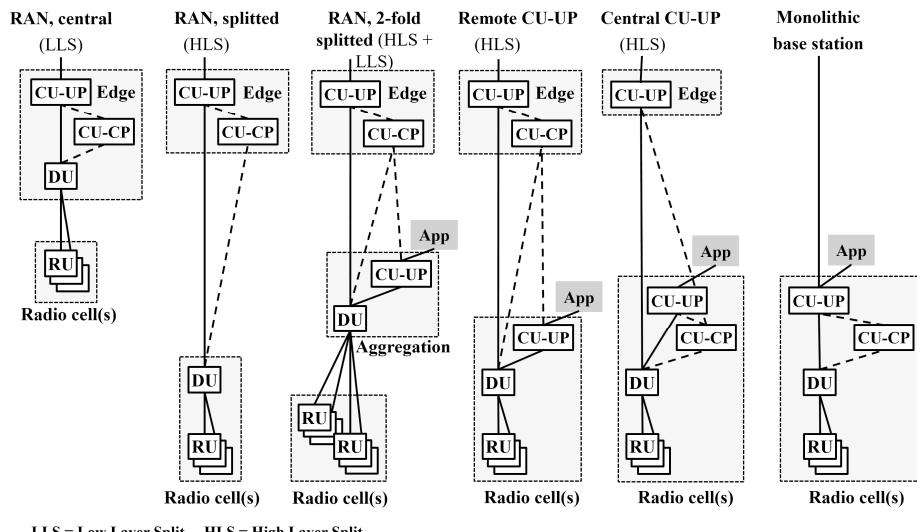


Fig. 7.12: Possible 5G-RAN architectures with function splitting [129]

The RAN architecture in Figure 7.11 also shows the reference points for the standardized interfaces, here – in contrast to Figures 6.1 and 6.2 – even with the differentia-

tion between control and user plane. the reference points listed below identify the interfaces to the core network and between the base stations [19; 129]:

- gNB – 5GC: NG
- gNB – gNB: Xn
- gNB – ng-eNB: Xn
- eNB – EPC: S1
- eNB – en-gNB: X2
- eNB – eNB: X2.

For the pure 5G reference points NG and Xn, the following figures 7.13 and 7.14 show the protocol stacks according to [51]. User and control plane are distinguished.

As shown in Figure 7.13, the user data packets are transmitted at the NG-U reference point tunneled via GTP-U (GPRS Tunneling Protocol-User plane) based on the connectionless UDP and IP between RAN and the 5GC function UPF (User Plane Function). Figure 7.13 also shows the protocol stack at the NG-C reference point for signaling using the NGAP (NG Application Protocol) between RAN and the 5GC function AMF (Access and Mobility Management Function). The transport protocol used is the connection-oriented, reliable SCTP (Stream Control Transmission Protocol). This interface is used for UE context and UE mobility management, paging to call a UE, session management, and the exchange of NAS messages between AMF in 5GC and UE [51].

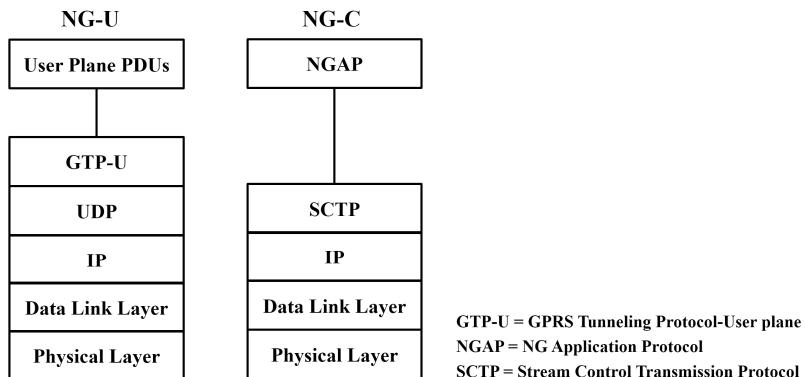


Fig. 7.13: NG protocol stacks between NG-RAN and 5G core [51]

The Xn interface enables direct communication between two NG-RAN nodes: gNBs, or ng-eNBs. Concerning the user data, the same protocol stack as in Figure 7.13 is used, as shown in Figure 7.14, i.e., GTP-U/UDP/IP. It also applies to the lower protocol layers for Xn-C compared to NG-C. Based on SCTP/IP, the gNBs are exchanging

XnAP messages (Xn Application Protocol). This interface is used for UE mobility management, UE context transfer, paging, and control of dual connectivity [51].

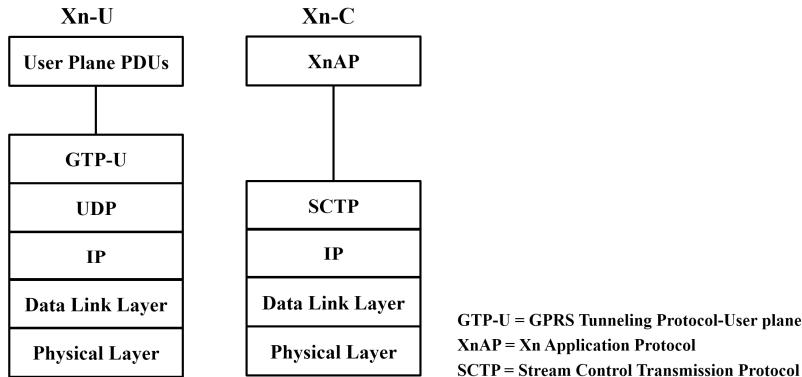


Fig. 7.14: Xn protocol stacks between gNBs in NG-RAN [51]

Concerning a 5G RAN, the standardization also provides for RAN sharing between two or more network operators. In other words, a RAN, including frequency resources, is shared, and the required resources are allocated [51; 37].

7.3 Open-RAN (O-RAN)

The O-RAN Alliance [150], an association of mobile network operators and manufacturers, pursues an exciting approach for a 5G-RAN. In addition to the 5G design principles outlined in Section 6.1 and applied to the RAN above, the O-RAN Alliance aims to create an open and intelligent 5G RAN as a basis for other, also smaller vendors and own developments by network operators. This shall be achieved by

- open, interoperable interfaces and APIs,
- the comprehensive use of virtualization, including the usage of the SDN concept with the introduction of a RIC (RAN Intelligent Controller),
- the use of artificial intelligence (AI) for automated network operation,
- open source software, and
- standard hardware

for the realization of a 5G base station.

Figure 7.15 shows the O-RAN reference architecture with the interfaces and functionalities from Figures 7.11 and 7.13 to 7.14. Also, other interfaces are specified by the O-RAN Alliance and a Service Management and Orchestration (SMO) system, including RICs [109; 209].

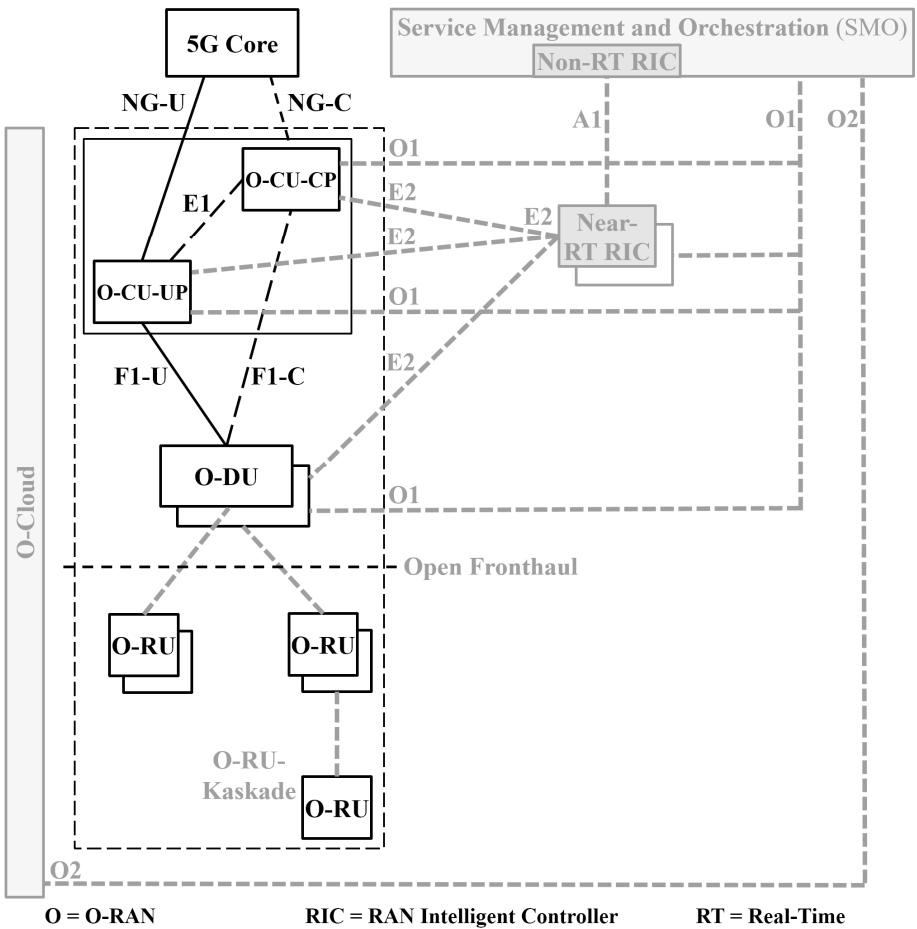


Fig. 7.15: O-RAN reference architecture [209]

The RAN interfaces NG-C, NG-U, E1, F1-C, and F1-U have been defined and standardized by 3GPP (see Section 7.2). New in the context of O-RAN are the additionally standardized RAN interfaces E2, O1, A1, O2, and, above all, an Open Fronthaul Interface between DU and RU. Compared with a 3GPP gNB solution, this provides open interfaces for the management and optimization of RAN functions via E2 and O1 and, in particular, the connection of RU technology via an open fronthaul interface. This enables not only the use of system technology from different manufacturers in the 5G core and access network but also of RAN transmission technology components from different manufacturers, especially DU and RU. The fronthaul interface between DU and RU was not standardized before the O-RAN activities. In addition, a more detailed standardized basis is created: for the optimization of gNB functions

during operation through the use of RICs (RAN Intelligent Controllers) and for the realization of a C-RAN (see Section 3.1) through the O2 interface to the O-Cloud.

Of course, the manifold function splitting shown for a gNB in Section 7.2, specifically in Figure 7.12, also applies to an O-RAN.

As already indicated, the SMO system is responsible for network management and orchestration of the O-RAN. To enable this, it provides the RAN-specific FCAPS (Fault Management, Configuration, Accounting, Performance and Security) functions required for this, the Non-RT RIC (Non-Real-Time RIC) for RAN optimization, and the mechanisms for orchestration, workflow, and O-cloud management. For this purpose, the following interfaces were specified: A1 between the non-RT RIC in SMO and the near-RT RIC for RAN optimization, O1 between SMO and the O-RAN network functions for FCAPS support, and O2 between SMO and the O-cloud for providing the necessary platform resources and their management.

The main goal of Non-RT RIC, with typical execution times of 1 s and more, is to support intelligent RAN optimization by providing policy-based defaults, machine learning models, and additional information through special applications called rApps (Non-RT RIC Application) for the Near-RT RIC function, so that the RAN can optimize, e.g., radio resource management. In addition, the Non-RT RIC can also perform intelligent radio resource management in non-real-time intervals (i.e., greater than 1 second). The non-RT RIC can use data analytics and AI/ML training to complement this.

The near-RT RIC provides a logical function that enables near-real-time control and optimization of the functions and resources of the RAN network elements (O-CU-CP, O-CU-UP, O-DU) controlled via E2 interfaces using fine-grained data acquisition and actions via the E2 interface with control loops in the order of 10 ms to 1 s, e.g., for automated radio network planning. For this purpose, the near-RT RIC hosts one or more xApps using the E2 interface to collect information in near real-time (e.g., on a UE or cell basis) and provide corresponding control and management services. The control of the RAN network elements via E2 is based on the policies and additional data provided by the non-RT RIC via A1.

Finally, the O-Cloud in Figure 7.15 is a cloud computing platform that includes the physical computing infrastructure to host the virtual O-RAN functions Near-RT RIC, O-CU-CP, O-CU-UP, and O-DU, as well as to provide the software components required for them, such as operating system, hypervisor, container runtime environment, etc., and the corresponding management and orchestration functions [209].

Based on the functions and optimization possibilities provided with an O-RAN, the O-RAN Alliance has specified and, to a large extent, already elaborated use cases according to the following list, which exemplifies how an O-RAN can be optimized during operation [210; 211]:

- Context-based dynamic handover management for V2X
- Flight path based dynamic UAV radio resource allocation

- Radio resource allocation for UAV application scenario
- QoE optimization
- Traffic steering
- Massive MIMO beamforming optimization
- RAN sharing between different operators
- QoS based resource optimization, e.g., for emergency services
- RAN Slice SLA assurance (Service Level Agreement) regarding bit rate, latency, availability, etc. It should be noted at this point that the topic of network slicing (see Section 8.4) is explicitly considered in O-RAN [212].
- Multi-vendor slices, i.e., DU and CU functions are from different vendors.
- Dynamic Spectrum Sharing (DSS) between 5G and 4G accesses
- Optimization of NSSI resource allocation (Network Slice Subnet Instance, see Section 8.4) regarding eMBB, URLLC, and mMTC
- Local indoor positioning in the RAN by using the corresponding xApp in the near-RT RIC
- Massive MIMO Single User/Multi User MIMO grouping optimization
- O-RAN signaling storm protection
- Congestion prediction and management for a base station
- Optimization for Industrial IoT (IIoT), including PDU session duplication, Time Sensitive Networking (TSN)
- BBU pooling to achieve RAN elasticity through flexible allocation of O-RUs to BBUs in the cloud (see Section 3.1)
- Integrated SON (Self-Organizing Networks) functions for automated configuration and optimization
- Shared O-RU between multiple O-DUs from one or more network operators
- Energy savings through efficient software and optimized configuration or control
- Multi User MIMO optimization.

Regarding O-RAN and security, we refer to Chapter 10.

In summary, O-RAN provides the following new opportunities: Until now, gNBs have been completely sourced from one manufacturer, mainly because of the non-open fronthaul interface. This restriction is dissolved by disaggregating a gNB up to the RU, and interoperability between different manufacturers is given. According to O-RAN, virtualization and the associated use of standard hardware is becoming the norm. However, the consequence is that compliance with the ambitious performance and energy efficiency requirements can be difficult. O-RAN pushes AI/ML-driven optimization and automated RAN operation through the RICs. The introduction of the SMO framework enables vendor-independent network operation of the RAN and its orchestration, as well as the use of open source software [122].

8 5G Core Network

It was already evident in Chapter 4 and Section 6.1, that a monolithic 5G system cannot fulfill the very different and sometimes extreme requirements for eMBB (Enhanced Mobile Broadband), URLLC (Ultra-Reliable and Low Latency Communications), and/or mMTC (Massive Machine Type Communications). This had a significant impact on the design principles for a 5G system outlined in Section 6.1. From this list, the following are particularly relevant for the core network: modularization, network softwarization, multi-tenant capability, cloudification, openness for 3rd-party providers, support of a wide range of wireless and wired access network technologies, and decoupling of the core network from the access network technology. Modularization plays an outstanding role here.

Figure 8.1 illustrates this in a first approximation. We can structure the overall 5G system in four areas:

- Terminal equipment (UEs)
- Access Network (AN) with various 3GPP and non-3GPP RATs and wired subscriber interfaces
- Core Network (CN)
- Data Network with the actual applications or communication services for the users.

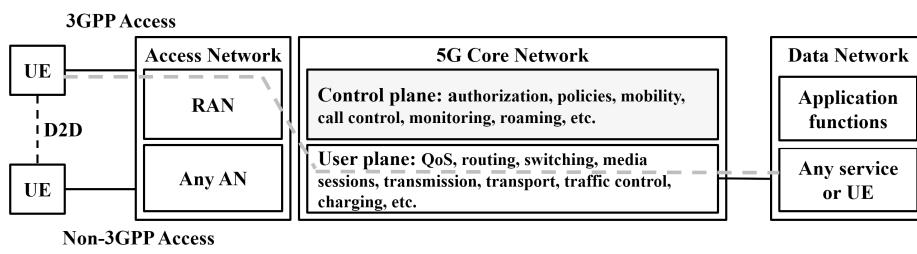


Fig. 8.1: Structuring of a 5G system

Figure 8.1 also shows the division of the network functions into those for the Control Plane (CP) with the signaling and control protocols and the User Plane (UP) for the user data, which is extremely helpful for modularization. It was already described for the 5G-RAN in Section 7.2. The UP includes functions such as user data transport, routing, and forwarding of data packets, traffic control, provisioning of the required QoS, ensuring service continuity for mobile use, and recording of billing data. The CP is responsible for authentication and authorization, compliance with the specified or agreed policies for users and the network, mobility management including

roaming, and the connection of 3rd party providers to the 5G core network concerning signaling, control, and monitoring.

The actual communication services are provided end-to-end via application functions, as shown in Figure 8.1. They use the CP and mainly the UP functions of the 5GC but are not part of it. One example is telephony, which is implemented by the complex application function IMS (see Sections 2.2 and 2.6).

The functions required in the UP and CP of the 5GC are provided as function modules. Sections 8.1 and 8.2 give an overview of these relatively fine-grained network functions (NF). These NFs are hosted in a repository and called via APIs. The underlying concept is called Service Based Architecture (SBA). Section 8.3 describes it in detail. According to the use cases to be supported and the associated requirements (see Chapter 4), NFs can then be composed and combined as required. It allows application, operator, and/or 3rd party provider-specific logical networks to be formed based on a 5G system. This technique is called Network Slicing and is explained in Section 8.4.

8.1 Basic System Architecture and Protocols

TS 23.501 [37] provides essential information on the system architecture of a 5G system and the required functions. Therefore, we start by summarizing the main principles and concepts to be applied:

- Separation of the UP and CP functions to ensure independent scalability, evolutionary development, and flexible applicability both at a central location and distributed in the network
- Modularization of the functional design to enable flexible and efficient network slicing
- Mapping processes, i.e., interactions between network functions, to services wherever possible, with the aim of easy reusability
- Where possible, network functions should communicate directly with each other, if necessary, via a proxy function.
- Minimizing the dependencies between the AN and the CN. The convergent CN should support different AN types and technologies.
- Provision of a uniform authentication system
- Support of “stateless” NFs, where the computing resources are decoupled from the memory resources
- Offering open interfaces for 3rd party users
- Support for simultaneous access to local and central NFs. For example, to offer very low latency services, UP functions are provided in the AN, the corresponding CP functions in the CN.
- Roaming with both routing of traffic on the home network and local forwarding on the visited network.

In the following, the 5G Core is described in several steps, with more details added. We start with the 5G basic system architecture, as shown in Figure 8.2. This shows the basic 5GC network functions UPF (User Plane Function) in the User Plane (UP) and AMF (Access and Mobility Management Function), and SMF (Session Management Function) in the Control Plane (CP). These three 5GC NFs are always required when user data is exchanged between a base station in the RAN and the Data Network (DN), e.g., the Internet, via the 5G core.

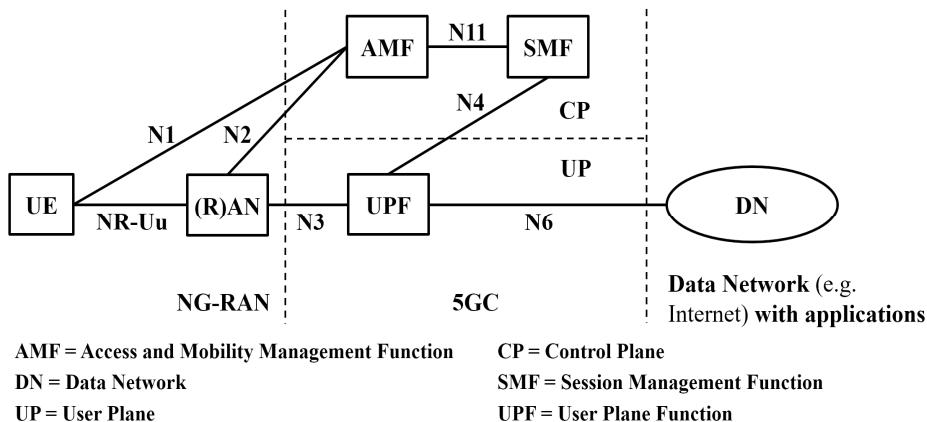


Fig. 8.2: 5G basic system architecture [37]

Figure 8.2 shows, in addition to the 5GC network functions AMF, SMF, and UPF mentioned above, the (R)AN with the UE terminal connected, and the external DN with the application functions used by the UE (e.g., web server for WWW services or IMS for telephony), standardized reference points Nx for their interconnection. All network functions shown in Figure 8.2 communicate peer-to-peer with each other: UE and gNB in the RAN via the NR-Uu interface, UE and AMF via N1, gNB and AMF via N2, gNB and UMF via N3, UPF and DN via N6, where the SMF controls the UPF via the N4 interface. N11 between AMF and SMF takes a special role since it is a service-based interface as part of the Service Based Architecture (SBA), as explained in more detail in Section 8.3 [37].

In the next step, the main functions of gNB, UPF, AMF, and SMF are explained before the mentioned interfaces are discussed.

The base station, a gNB or ng-eNB (see Section 7.2), provides the radio transmission technology to the UEs (see Section 7.1). This includes radio resource management functions such as Radio Bearer Control, Radio Admission Control and Connection Mobility Control, and the dynamic allocation of transmission resources to the UEs in the uplink and downlink directions (scheduling). The headers of the IP packets or Ethernet frames transmitted over the air interface are compressed, and the

data is encrypted. If necessary, an AMF is selected. CP messages are forwarded to the AMF; UP data is sent to the UPF. In the direction of the UE, connections are established and terminated, and paging or broadcast messages are transmitted. For QoS, a gNB supports the management of the flows, mapping to the radio channels, and marking data packets. Network slicing may be supported. In addition, RAN sharing, i.e., the common use of frequencies by two base stations or the use of one gNB by two providers, and dual connectivity, i.e., the connection of one UE to two gNBs, are supported. In addition, transmission-related measurements are performed, and the results are provided [51].

The UPF (User Plane Function) is responsible for the user data handling, the PDU sessions (Protocol Data Unit), in the 5GC. It represents the anchor point for traffic via N3 (NG-U) from and to the AN for mobile use and the transition to the DN via N6. Its tasks naturally include routing and forwarding of data packets, including traffic control and rerouting, in addition to QoS handling (including ensuring DL and UL bit rates, marking packets, assigning packets to flows), packet inspection, lawful interception (LI), and collecting and providing usage data. UPF behavior is controlled and programmed by the SMF via the N4 reference point [37; 51].

The AMF (Access and Mobility Management Function) in the CP terminates the NAS session signaling (Non Access Stratum protocol), transmitted transparently through the RAN, with the UE at the N1 reference point, and the signaling (NG-C) with the (R)AN at the N2 reference point (see section 7.2). Registration, connection, reachability, and mobility management, including authentication and authorization, also take place here. The SMF responsible for the desired PDU session is selected. The AMF ensures the reachability of the UEs in inactive mode, among other things, using so-called paging. In addition, the AMF provides security functions for the NAS (Non Access Stratum, communication between UE and 5GC) and the AS (Access Stratum, communication between UE and gNB) and supports network slicing. In addition, lawful interception and localization data are collected here [37; 51].

The SMF (Session Management Function) is mainly responsible for session control, i.e., establishing, modifying, and terminating PDU sessions, UE IP address management, and UPF selection and control via the N4 reference point. In this context, the routing in the UPF is configured, and the settings for implementing the QoS required in each case are determined. In addition, data for billing and also for LI is collected here [37; 51].

In the following, the protocol stacks of the control plane, i.e., at the interfaces NR-Uu, N2, N1, and N11, are discussed according to Figure 8.3. Concerning N11, please also refer to Section 8.3.

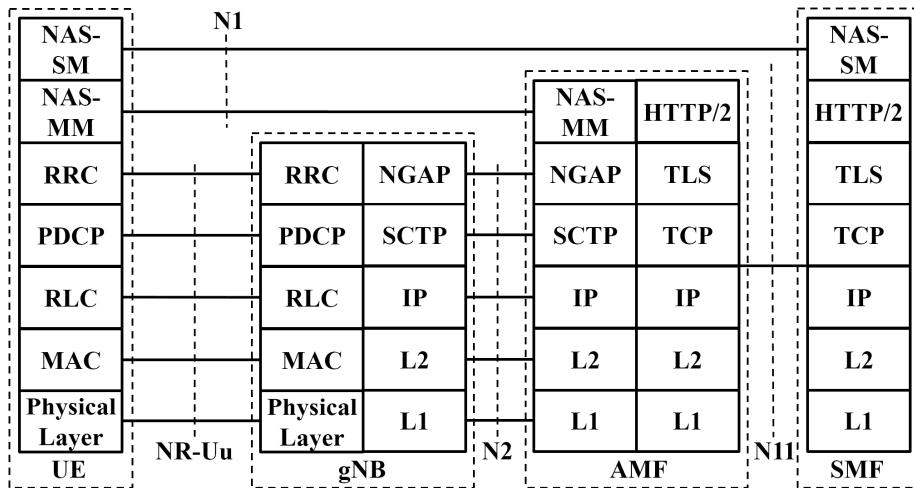


Fig. 8.3: Control plane protocol stacks

The protocol stack at the NR-Uu interface is based on the actual radio interface (Physical Layer) and the layer 2 MAC protocol. On top of this, the RLC protocol (Radio Link Control) transmits the higher layers' protocol messages. It provides sequence numbering, error correction with ARQ (Automatic Repeat Request), segmentation, detection of message duplicates, and protocol error detection [51]. The PDCP (Packet Data Convergence Protocol), which is based on RLC, transports both the signaling messages (see Figure 8.3) and the user data (see Figure 8.9). Functions for this also include sequence numbering, header compression, encryption, integrity protection, duplicate detection, and ensuring correct message order [51]. Finally, the Radio Resource Control (RRC) protocol uses the underlying protocols in the stack, PDCP/RLC/MAC/Physical Layer, to transmit AS and NAS broadcast system information, send paging messages, and establish, maintain, and re-enable RRC connectivity between the UE and gNB, including managing frequency carrier aggregation and dual connectivity. In addition, the RRC protocol provides data channels on the radio link for signaling and user data transmission. It ensures mobility management, including handover and radio cell selection by the UE. In addition, it supports the assurance of the required QoS, the handling of a radio link interruption, security functions including key management, measurements between UE and gNB, and, last but not least, the transport of the NAS signaling messages according to Figure 8.3 [51].

The N2 reference point between the RAN and AMF in Figures 8.2 and 8.3 corresponds to the NG-C interface in Figure 7.11. According to Figure 8.3, this CP protocol stack is based on IP in Layer 3. Layer 4 uses the reliable SCTP (Stream Control Transmission Protocol) according to RFC 4960 [214], which supports multiple simultaneous message flows. This, in turn, transports the NGAP (NG Application Protocol) signaling messages between the NG-RAN and the AMF. This provides, among other things, paging, UE context management, handover, RAN-specific PDU session management, the transmission of NAS signaling messages, and UE location information [215; 216; 37].

The interrelationships at N2 are further clarified below by an example from network practice. Figure 8.4 shows an NGAP message captured at N2 between gNB (in the practical example with the IP address 10.0.1.27) and AMF (IP address 10.0.1.20). Figure 8.5 shows a NAS-MM message transmitted to the AMF using NGAP.

```
Internet Protocol Version 4, Src: gNB-N2 (10.0.1.27), Dst: AMF (10.0.1.20)
Stream Control Transmission Protocol, Src Port: 53339 (53339), Dst Port: 38412 (38412)
NG Application Protocol (NGSetupRequest)
  NGAP-PDU: initiatingMessage (0)
    initiatingMessage
      procedureCode: id-NGSetup (21)
      criticality: reject (0)
    value
      NGSetupRequest
```

Fig. 8.4: NGAP message captured with protocol analysis software at interface N2

```
Internet Protocol Version 4, Src: gNB-N2 (10.0.1.27), Dst: AMF (10.0.1.20)
Stream Control Transmission Protocol, Src Port: 53339 (53339), Dst Port: 38412 (38412)
NG Application Protocol (InitialUEMessage)
  ▼ NGAP-PDU: initiatingMessage (0)
    ▼ initiatingMessage
      procedureCode: id-InitialUEMessage (15)
      criticality: ignore (1)
    ▼ value
      ▼ InitialUEMessage
        ▼ protocolIEs: 5 items
          ▷ Item 0: id-RAN-UE-NGAP-ID
          ▷ Item 1: id-NAS-PDU
            ▼ ProtocolIE-Field
              id: id-NAS-PDU (38)
              criticality: reject (0)
            ▼ value
              ▼ NAS-PDU: 7e004179000d0100f11000000000000000000102e04f0f0f0f0
                ▼ Non-Access-Stratum 5GS (NAS)PDU
                  ▼ Plain NAS 5GS Message
                    Extended protocol discriminator: 5G mobility management messages (126)
          •
          •
          •
```

Fig. 8.5: NAS MM message captured with protocol analysis software at interface N2

The N1 reference point between UE and AMF or SMF includes, as shown in Figure 8.3, the NAS (Non Access Stratum) signaling, i.e., signaling not related to the RAN. For each connected UE, one NAS signaling connection is terminated in the AMF. Between the UE and the AMF, the NAS-MM protocol (mobility management) is used at N1. It supports UE registration, connection management, and activation and deactivation of user data connections in the user plane. It is also responsible for encryption and integrity assurance of NAS signaling [37]. The NAS-SM protocol (session management) between UE and SMF at N1 supports session management between UE and the SMF. This is used to establish, modify and terminate PDU sessions in the UP via SMF. NAS-SM signaling messages are relayed transparently by the AMF (relay function). In addition to the NAS-SM messages between UE and SMF, the NAS-MM protocol is also used to transmit policies from the PCF (Policy Control Function, see Section 8.3) to the UE, to exchange SMS text messages between UE and SMSF (Short Message Service Function) and to provide localization information from the UE to the LMF (Location Management Function) [37].

The N11 reference point in Figure 8.3 between AMF and SMF, unlike the other reference points discussed in this section, does not represent an interface for peer-to-peer communication but an API based on HTTP/2 for exchanging data with all other relevant NFs in the Service Based Architecture (see Section 8.3) [43].

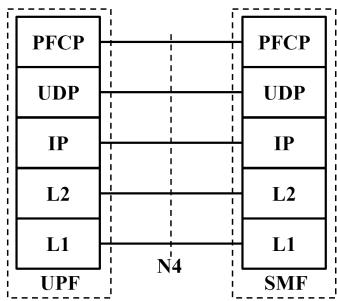
An example from network practice for N11 is shown in Figure 8.6 with a NAS-SM message captured between AMF (in the practical example with IP address 10.0.1.20) and SMF (IP address 10.0.1.25), where HTTP/2 over TCP is used here, i.e., transmitted unencrypted for illustrative reasons.

```

Internet Protocol Version 4, Src: AMF (10.0.1.20), Dst: SMF (10.0.1.25)
Transmission Control Protocol, Src Port: 59888, Dst Port: 7777, Seq: 216, Ack: 16, Len: 1133
HyperText Transfer Protocol 2
  > Stream: DATA, Stream ID: 1, Length 1124
  ▼ MIME Multipart Media Encapsulation, Type: multipart/related, Boundary: "=-FXv8V2Y6XN+j03hVa52Wzg=="
    [Type: multipart/related]
    First boundary: ---FXv8V2Y6XN+j03hVa52Wzg==\r\n
    ▶ Encapsulated multipart part: (application/json)
      Boundary: \r\n--=-FXv8V2Y6XN+j03hVa52Wzg==\r\n
    ▼ Encapsulated multipart part: (application/vnd.3gpp.5gnas)
      Content-Id: 5gnas-sm\r\n
      Content-Type: application/vnd.3gpp.5gnas\r\n\r\n\r\n
    ▼ Non-Access-Stratum 5GS (NAS)PDU
      ▼ Plain NAS 5GS Message
        • Extended protocol discriminator: 5G session management messages (46)
        •
        •
  
```

Fig. 8.6: NAS SM message captured with protocol analysis software at interface N11

As mentioned, an SMF controls the corresponding UPF(s) via the N4 interface. Figure 8.7 shows the protocol stack [213].



PFCP = Packet Forwarding Control Protocol

Fig. 8.7: N4 Interface with PFCP [213]

The crucial protocol at N4 is the PFCP (Packet Forwarding Control Protocol). It uses UDP and IPv4 or IPv6 to transport the PFCP messages. Layers 2 and 1 are typically based on an Ethernet interface, but other L2 and L1 protocols or interfaces can also be used. PFCP can be compared, in a first approximation, to OpenFlow or similar protocols in an SDN (see Section 3.2). It provides the signaling to control the user plane, the UPF, from the control plane, the SMF, and thus to control the handling of the user data packets in the UP according to the service requirements. However, the PFCP is tailored to the specific conditions of the mobile network. In this respect, unlike an SDN controller, the SMF knows the required handling of user data packets arriving at the UPF because of the PDU sessions established by signaling. Using the PFCP standardized in TS 29.244 [213], the SMF controls the packet processing in the UPF per PDU session. For this purpose, PFCP session contexts are created, modified, or deleted. Within such a session context, so-called PDRs (Packet Detection Rule), FARs (Forwarding Action Rule), QERs (QoS Enforcement Rule), URRs (Usage Reporting Rule), BAR (Buffering Action Rule), and/or MAR (Multi-Access Rule) are added, modified or deleted, or predefined PDRs, FARs, QERs, URRs are enabled or disabled. Each PDR contains a PDI (Packet Detection Information), i.e., one or more match fields, based on which incoming packets in the UPF are checked and, in case of a match, processed according to the instructions from at least one or several corresponding FARs. Depending on the PDR, instructions from QERs, URRs, or a MAR can also be executed. A FAR can use the Action Apply parameter, for example, to initiate the forwarding, duplication, or dropping of a packet or to accept or reject the request to join an IP multicast group. If there is no match with a PDR, the UPF drops the packet. In addition to the described function, the N4 interface can also be used to transport UP information transmitted tunneled with GTP-U (GPRS Tunnelling Protocol-User plane). Examples are DHCP (Dynamic Host Configuration Protocol) messages from the SMF to a UE for IP address assignment or AAA information from a server in the DN to the SMF [213].

An example from network practice for N4 is shown in Figure 8.8, with a PFCP message captured between SMF (in the practical example with IP address 10.0.1.25) and UPF (IP address 10.0.1.26). The corresponding PDR causes, among other things, a packet with source *Access*, and destination (Network Instance) *Internet*, to be forwarded according to the PDR and the associated FAR.

```

Internet Protocol Version 4, Src: SMF (10.0.1.25), Dst: UPF-N4 (10.0.1.26)
User Datagram Protocol, Src Port: 8805, Dst Port: 8805
Packet Forwarding Control Protocol
  > Flags: 0x21, SEID (S)
    Message Type: PFCP Session Establishment Request (50)
    Length: 513
    SEID: 0x0000000000000000
    Sequence Number: 5
    Spare: 0
  > Node ID : IPv4 address: 10.0.1.25
  > F-SEID : SEID: 0x0000000000000001, IPv4 10.0.1.25
  > Create PDR : [Grouped IE]: PDR ID: 1
  < Create PDR : [Grouped IE]: PDR ID: 2
    IE Type: Create PDR (1)
    IE Length: 69
    > PDR ID : 2
    > Precedence : 255
    < PDI : [Grouped IE]
      IE Type: PDI (2)
      IE Length: 29
      > Source Interface : Access
      > F-TEID :
      > Network Instance : internet
      > QFI :
      > Outer Header Removal : GTP-U/UDP/IPV4
      > FAR ID : Dynamic by CP 2
      > QER ID : Dynamic by CP 1
    > Create PDR : [Grouped IE]: PDR ID: 3
    > Create PDR : [Grouped IE]: PDR ID: 4
    > Create FAR : [Grouped IE]: FAR ID: Dynamic by CP 1
    < Create FAR : [Grouped IE]: FAR ID: Dynamic by CP 2
      IE Type: Create FAR (3)
      IE Length: 22
      > FAR ID : Dynamic by CP 2
    < Apply Action :
      IE Type: Apply Action (44)
      IE Length: 1
      0... .... = DFRT (Duplicate for Redundant Transmission): False
      .0... .... = IPMD (IP Multicast Deny): False
      ..0. .... = IPMA (IP Multicast Accept): False
      ...0 .... = DUPL (Duplicate): False
      .... 0... = NOCP (Notify the CP function): False
      .... .0.. = BUFF (Buffer): False
      .... ..1. = FORW (Forward): True
      .... ...0 = DROP (Drop): False
    > Forwarding Parameters : [Grouped IE]
  > Create FAR : [Grouped IE]: FAR ID: Dynamic by CP 3
  > Create URR : [Grouped IE]: URR ID: Dynamic by CP 1
  > Create QER : [Grouped IE]: QER ID: Dynamic by CP 1
  > Create BAR : [Grouped IE]: BAR ID: 1
  > PDN Type : IPv4

```

Fig. 8.8: PFCP message captured with protocol analysis software at interface N4

Having discussed the CP functions and protocols in a 5G system above, we now look closer at the UP protocol stacks, i.e., the NR-Uu, N3, and N6 interfaces, as shown in Figure 8.9.

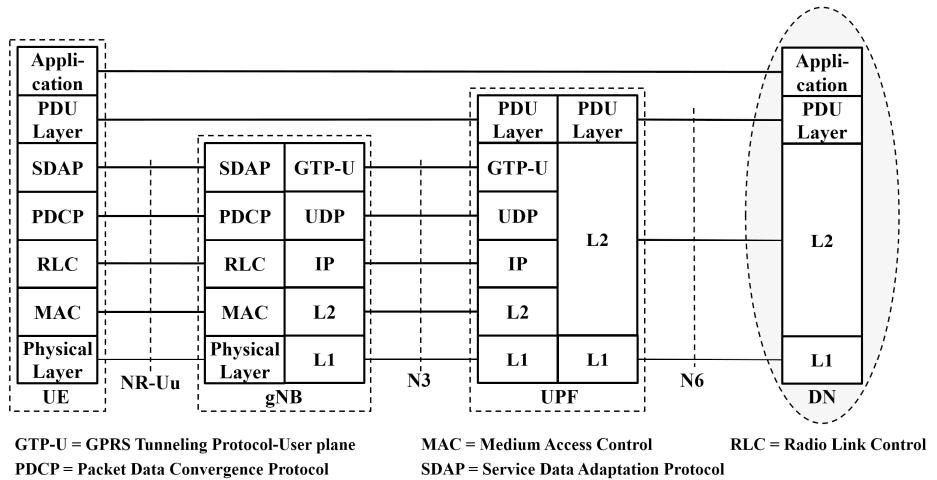


Fig. 8.9: User plane protocol stack [51; 37]

The UP protocol stack at the NR-Uu interface is the same as that for the CP, except that the SDAP (Service Data Adaptation Protocol) is now used instead of the RRC protocol: SDAP/PDCP/RLC/MAC/Physical Layer. The main functions of SDAP are mapping a QoS flow and the associated data channel on the radio link (data radio bearer) in both downlink and uplink directions. In this context, a QoS flow of a PDU session means all user data packets with the same QoS handling. All packets assigned to the same QoS flow get the same forwarding processing, e.g., the same policies for scheduling, queue management, rate shaping, RLC configuration, etc. A QoS flow is identified by a QoS Flow ID (QFI) [51].

The N3 reference point between RAN and UPF in Figures 8.2 and 8.9 corresponds to the NG-U interface in Figure 7.11. According to Figure 8.9, this UP protocol stack is based on IP in Layer 3. In Layer 4, connectionless UDP is used concerning possible real-time user data. The user data to be transmitted is transported in tunnels from the (R)AN or gNB to the UPF at the interface to the data network using the GTP-U (GPRS Tunnelling Protocol-User plane). I.e., the protocol specified in TS 29.281 [217] provides one tunnel through the 5G system per PDU session or QoS flow. The advantage of this tunneling is that the actual payload packets between UE and UPF or DN can use the same and constant IP addresses per PDU session, even if, for example, the gNB and thus its IP address at the N3 interface changes as a result of mobility. The tunnel endpoint in the gNB is adapted under SMF control. Handover

or roaming between different gNBs is thus possible if the PDU session is still active [37].

Figure 8.10 shows an example from network practice for N3 with an ICMP/IPv4 packet captured between gNB (IP address 10.0.2.20 in the practical example) and UPF (IP address 10.0.2.21). The IP addresses mentioned are the GTP-U tunnel endpoints (underlay network). The IP addresses used for the ICMP application for end-to-end communication between UE and server in the DN (Overlay Network) are 10.45.0.2 (UE) and 10.0.3.21 (server).

```

Internet Protocol Version 4, Src: gNB-N3 (10.0.2.20), Dst: UPF-N3 (10.0.2.21)
User Datagram Protocol, Src Port: 2152, Dst Port: 2152
GPRS Tunneling Protocol
  > Flags: 0x34
    Message Type: T-PDU (0xff)
    Length: 92
    TEID: 0x00000002 (2)
    Next extension header type: PDU Session container (0x85)
  ✓ Extension header (PDU Session container)
    Extension Header Length: 1
    ✓ PDU Session Container
      0001 .... = PDU Type: UL PDU SESSION INFORMATION (1)
      .... 0000 = Spare: 0x0
      00... .... = Spare: 0x0
      ..00 0001 = QoS Flow Identifier (QFI): 1
    Next extension header type: No more extension headers (0x00)
Internet Protocol Version 4, Src: UE (10.45.0.2), Dst: DN (10.0.3.21)
Internet Control Message Protocol

```

Fig. 8.10: ICMP message captured with protocol analysis software at interface N3

End-to-end PDU sessions are established between UE and UPF for the user data, according to Figure 8.9. These can be of type IPv4, IPv6, IPv4 and IPv6, Ethernet, or even Unstructured. In the first three cases, the PDU sessions or the associated IP packets are identified by IP addresses; in the fourth case, the transmitted Ethernet frames are identified by MAC addresses. In the latter case, the user data can be transported using any protocol, e.g., in the context of IoT applications [37].

At reference point N6, the UPF transfers the user data of a PDU session to the external data network or accepts it from there. Here, the protocol stack used in the DN for the application is applied, e.g., ICMP/IPv4/L2/L1.

Figure 8.11 shows an ICMP message from network practice captured at N6.

```

Internet Protocol Version 4, Src: UPF-N6 (10.0.3.20), Dst: DN (10.0.3.21)
Internet Control Message Protocol
Type: 8 (Echo (ping) request)

```

Fig. 8.11: ICMP message captured with protocol analysis software at interface N6

8.2 Core Network Functions

Section 8.1, especially Figure 8.2, describes the 5G base system architecture. The separation between UP and CP functions, the minimization of dependencies between the access network (AN) and core network (CN), and the modularization became obvious. The latter is even more significant than previously presented and, therefore, elaborated in this and the following section. The fine-granular modularization carried out in a 5G system and explained below by introducing small-scale, cooperating network functions (NFs) ensures that the principles and concepts essential for 5G mentioned at the beginning of Section 8.1 could all be implemented.

The overall system architecture for a 5G system is described in TS 23.501 [37], focusing on the CN. The CN network functions UPF, AMF, and SMF, which are most important for the basic function, have already been explained above in Section 8.1, and their interaction has also been described. Based on Figure 8.12, we discuss now the other NFs in the CN. In addition, the UPF is examined in more detail.

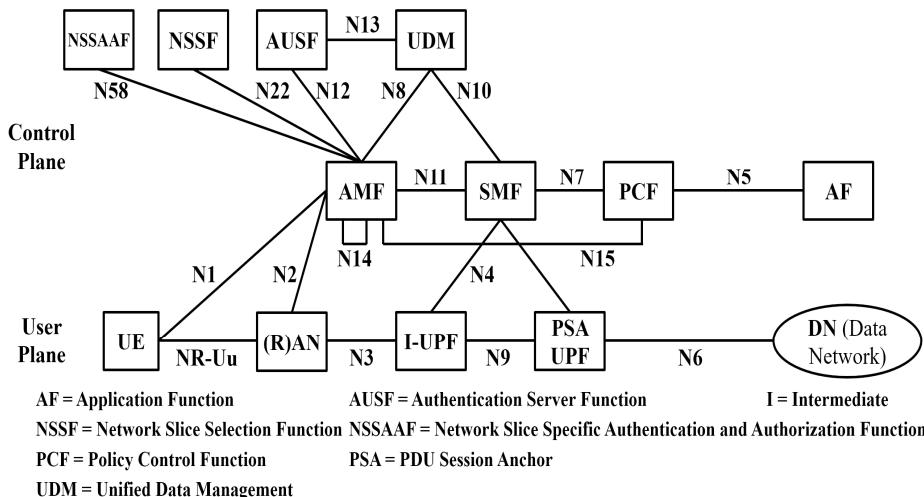


Fig. 8.12: Network functions with reference points for a 5G system without roaming [37]

Figure 8.12 shows the most important network functions listed in [37] in a representation with reference points Nx, i.e., these NFs communicate peer-to-peer via standardized reference points, which in the control plane are APIs (see Section 8.3). As already explained in Section 8.1, the AMF (Access and Mobility Management Function) is primarily the first point of contact for the UE, and the (R)AN for signaling, selects the SMF for a PDU session, and provides functions for registration, connection, reachability and mobility management including authentication and authorization. The SMF (Session Management Function) is essentially responsible for session control, i.e., setting up, modifying, and terminating PDU sessions, IP address management, UPF selection and control via the N4 reference point, and information exchange with the PCF (Policy Control Function) via N7. The UPF (User Plane Function) is responsible for the user data handling, the PDU sessions (Protocol Data Unit), in the 5GC. It represents the anchor point for traffic via N3 from and to the AN for mobile use and the transition to the DN via N6. Its tasks naturally include routing and forwarding data packets, including traffic control and rerouting, and QoS handling. The UPF behavior is controlled and programmed by the SMF via the reference point N4 [51].

Regarding the UPF, two UPFs connected via reference point N9 are shown in Figure 8.12, an I-UPF (Intermediate UPF) and a PSA UPF (PDU Session Anchor). The PSA UPF must always exist in a 5GC. It implements the IP anchor, i.e., the GTP-U tunnel endpoint at the transition to the DN, and is thus essential for mobility. Even if the UE moves and, as a result, the gNB and, thus, its IP address changes, the tunnel through the 5GC remains unchanged for the application data. In addition, an I-UPF can be connected upstream of the PSA UPF. Since such an I-UPF can be placed anywhere in the network, e.g., in the edge area at the location of a base station, the user data can be handled in the I-UPF according to the application requirements. E.g. for a V2X application with challenging delay requirements, the user data could already be routed at the gNB location. The possibility to place a UPF in the network independently from the CP-NFs like AMF and SMF is an essential requirement for URLLC applications with very short delay times. In addition, there can be multiple I-UPFs and PSA UPFs to provide redundant paths for PDU sessions, in extreme cases, from the UE via two gNBs through the 5G core to the DN. This makes extremely high reliability and availability feasible in the URLLC case, especially since multiple SMFs can also be used. If, instead, extensive sensor data is to be collected centrally and transferred to the application in the DN in an mMTC use case, it makes sense to provide a central UPF in the cloud at the transition to the DN. For a deeper understanding of a network situation with I-UPF and PSA UPF, Figure 8.13 shows the UP protocol stacks, specifically the protocols at the N9 interface (see Section 8.1., Figure 8.9) [37].

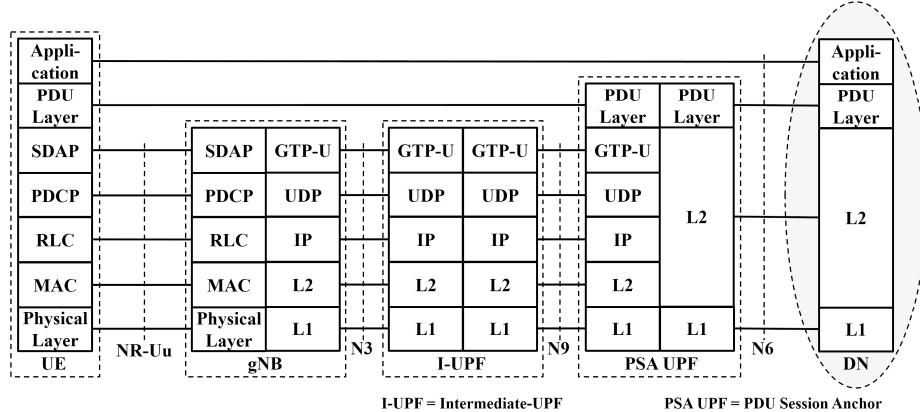


Fig. 8.13: User plane protocol stacks with I-UPF and PSA UPF [37]

Returning to the CP-CN in Figure 8.12, it is clear that in addition to AMF and SMF, numerous other NFs are essential for the function of a 5G system. First, the PCF (Policy Control Function) should be mentioned here. It makes uniform policies for network behavior (e.g., QoS, traffic forwarding, AN priorities) and makes them available to other NFs. To do this, it accesses the user profiles and application requirements in the UDR (Unified Data Repository). The UDM (Unified Data Management) handles authentication and identification data using the user profiles (subscription data), and the AUSF (Authentication Server Function) represents an authentication server. In the case of network slicing, the NSSF (Network Slice Selection Function) selects the network slice(s) responsible for a UE, manages NSSAI (Network Slice Selection Assistance Information) and subscribed S-NSSAIs (Single-NSSAIs), and determines the AMF instance(s) for them. The NSSAAF (Network Slice Specific Authentication and Authorization Function) provides slice-specific authentication and authorization. Finally, the AF (Application Function) represents an additional CP function that represents an application used over the 5G network and can also be offered by a 3rd party provider. It is not part of the 5GC but can communicate with the NFs of the 5GC, preferably via a NEF (Network Exposure Function). The 5G network operator may allow a trusted AF to interact directly with the relevant NFs in the 5GC. This can go so far that an AF can influence the policies applied and the routing decisions of the SMF to PDU sessions, including UPF selection [37].

The CP-NFs discussed so far and shown in Figure 8.12 communicate according to the reference points mentioned above, firstly peer-to-peer, i.e., directly, with each other, and secondly via APIs. In addition, a whole range of network functions communicates with the NFs already briefly explained and with each other only via APIs. All of these NFs, which interact with the same protocol stack via RESTful APIs, are

part of the Service Based Architecture (SBA), which is described in more detail in the following section, with the APIs summarized under the term Service-based Interfaces (SBI, see Figure 8.15). Before discussing the SBA, however, an overview of the NFs in the 5GC that have not yet been discussed should be given.

The NEF (Network Exposure Function) has already been mentioned above for the secure connection of an AF to the 5GC. It announces 5G internally available functionalities, resources, and events to 3rd party providers, AFs, or edge computing applications. In addition, the NEF ensures the secure provisioning of the application's data in the 5GC. For this purpose, the NEF can authenticate and authorize the AF. In addition, it translates between 5G-internal and AF-external information, e.g., it can map an AF service identifier to 5GC information such as DNN (Data Network Name) and S-NSSAI (Single-Network Slice Selection Assistance Information). It can also provide PFDs (Packet Flow Descriptions) for the SMF or manage 5G LAN groups. Obviously, it collects charging data and makes it available again. The NEF uses the UDR (Unified Data Repository) for structured data storage. Here, it stores, for example, data from other NFs and makes them available to an AF [37].

A crucial network function in 5GC is the NRF (Network Repository Function). All available NFs or instances with their NF profiles and network services are known to it and can be queried there by other NFs that want to use their services (discovery). It notifies about newly registered, updated, or deregistered NF instances. In the context of network slicing (see Section 8.4), multiple NRFS can be deployed across and/or within slices [37].

The UDR (Unified Data Repository) for structured data storage has already been mentioned, for example, in the context of the NEF. In addition, it serves the UDM for storing user data, the PCF for policies, the AFs for PFDs, etc. The UDSF (Unstructured Data Storage Function) is available for unstructured data. The SCP (Service Communication Proxy) provides routing and message forwarding to target NFs, security functions, aggregation of messages from NFs, load balancing, and congestion control for scalable communication in the SBA. Its function in the SBA is reminiscent of the STP (Signaling Transfer Point) for No. 7 signaling in 2G and 3G networks and the DSC (Diameter Signaling Controller) for Diameter messages in 4G networks. The 5G EIR (Equipment Identity Register) can be used to check the status of the PEI (Permanent Equipment Identifier) of terminal equipment, e.g., to prevent the use of a stolen UE in the event of a PEI blacklist status. The UCMF (UE radio Capability Management Function) stores the radio characteristics of UEs to optimize the corresponding signaling. The NWDAF (Network Data Analytics Function) collects data from NFs, AFs, and management for network analysis and makes it available to NFs and AFs. The Charging Function (CHF) supports collecting and providing charging information. The SEPP (Security Edge Protection Proxy) is a non-transparent proxy that monitors and filters signaling messages between the 5GCs of different providers regarding security and performs topology hiding, i.e., shields its

5GC from the outside world. Finally, the SMSF (Short Message Service Function) ensures the forwarding of SMS (Short Message Service) messages [37].

Before going into more detail on the interaction of the essential network functions described above, the SBA concept is explained in Section 8.3 for better understanding. But first, Figure 8.14 gives a more comprehensive overview of a 5G system, not only about network functions in UP and CP but also about the mandatory network management and orchestration functions. Concerning implementation, it is assumed here that the management functions are also integrated into the SBA.

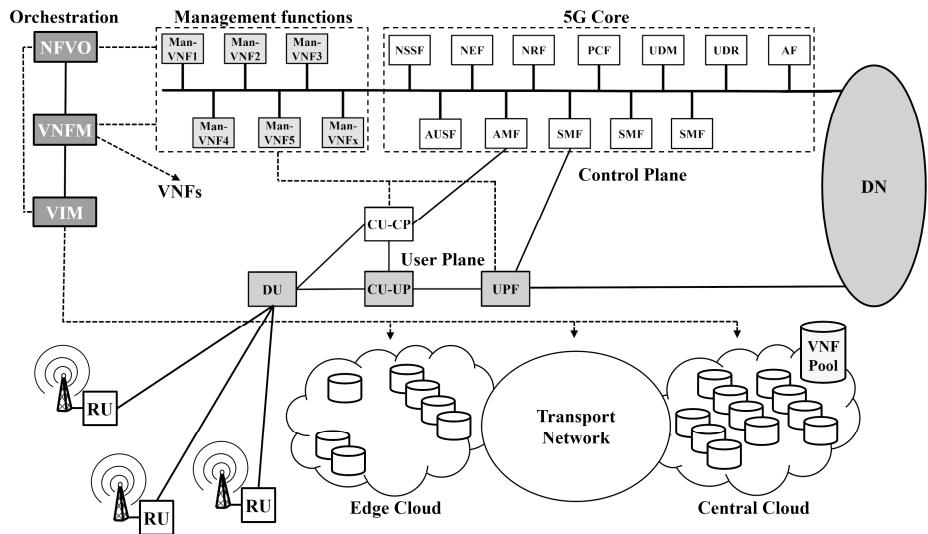


Fig. 8.14: Network functions for the user plane, control plan, network management, and orchestration in a 5G system

8.3 Service Based Architecture (SBA)

In conventional telecommunications networks, the network elements usually communicate point-to-point via specially specified interfaces, often with different transport protocol stacks. This led to relatively high complexity and effort for modifications and a poor reuse rate, in summary, to low flexibility and openness. These disadvantages should be avoided in a 5G system:

- Communication between network functions is possible in a flexible and dynamically changeable way, without a fixed point-to-point relationship.
- All network services are using a uniform protocol stack.
- A network function can make its services available to other network functions.
- Several versions of the same service can coexist simultaneously.

- A network function only has to take care of the services it offers and the services it uses and does not influence other services.
- All operations that concern the same communication context and can thus change are handled by one service.

This is achieved by:

- Each network function provides its services via an API.
- A network function registers with a central repository function with the services it offers.
- A network function requests a specific service in the repository on demand.

If this is guaranteed, each network function can have its life cycle, and new services can easily be introduced. It is also easy to reuse an existing service in a new application. Such a highly flexible and modular system represents a Service Based Architecture (SBA) [88].

Figure 8.15 shows the 5GC system architecture with the APIs already mentioned for the interaction of the NFs. This architecture view is the counterpart to the representation with the reference points in Figure 8.12. The APIs are represented by so-called Service-based Interfaces (SBI). The following SBIs were standardized in [37], whereby the NF providing the respective API is indicated in brackets: Namf (AMF), Nsmf (SMF), Nnef (NEF), Npcf (PCF), Nudm (UDM), Naf (AF), Nnrf (NRF), NnssAAF (NSSAAF), Nnssf (NSSF), Nauf (AUSF), Nudr (UDR), Nuds (UDSF), N5g-eir (5G-EIR), Nnwda (NWDAF), Nchf (CHF), Nucmf (UCMF).

For completeness, Figure 8.15 also shows the non-SBI reference points, i.e., the interfaces not implemented as APIs. The NFs communicating point-to-point via the reference point Nx are listed in brackets: N1 (UE – AMF), N2 ((R)AN – AMF), N3 ((R)AN – UPF), N4 (SMF – UPF), N6 (UPF – DN), N9 (UPF – UPF) [37].

An NF offers its services via its SBI, e.g., the NRF via Nnrf. Another NF can then use the services via this API, e.g., an AMF or SMF. The service provider is called producer; the service user is called consumer.

Accordingly, SBA supports the following mechanisms:

- A service of an NF is registered or deregistered. This provides the NRF (Network Repository Function) with information on all available NF instances and the services they offer.
- An NF can, therefore, request a required service from the NRF (service discovery).
- Every service usage must be authorized. The necessary authorization data are stored in the NF, offering the service [37; 43].

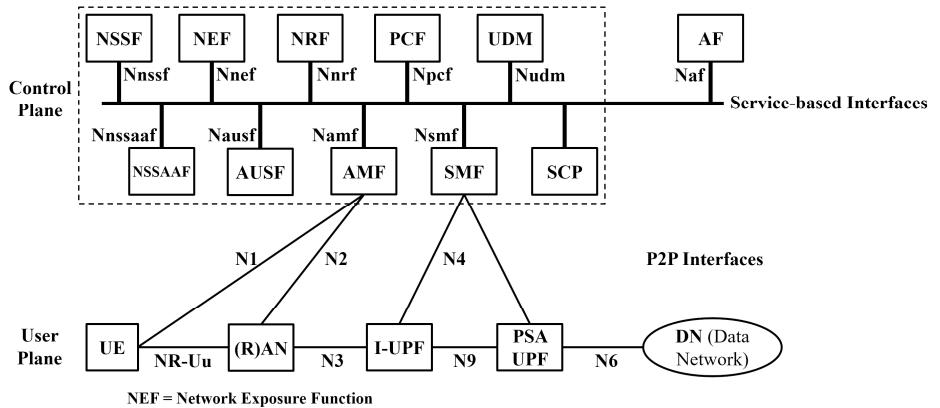


Fig. 8.15: Network functions with SBA interfaces for a 5G system without roaming [37]

An example illustrates the interaction of selected NFs from Figure 8.15. The PDU session setup initiated by a UE was chosen for this, i.e., the interaction of NFs to provide a path with defined QoS for the user data through the 5G system. Involved are the 5GC NFs AMF, SMF, UDM, PCF, and UPF. The sequence for this, which has been simplified for clarity, is shown in Figure 8.16.

It is crucial to know in advance that the SMF (Session Management Function) is responsible for the complete signaling and control of the UP functions within a PDU session. In this regard, the SMF has the following specific tasks:

- Selecting the UPF
- UPF control via N4
- Signaling via AMF with the (R)AN via N2 to exchange the QoS parameters
- Signaling via AMF with the UE via N1 to establish and terminate the PDU session and to transfer the QoS rules to the UE
- Communication with the AMF regarding signaling via N1 and N2. Besides, the SMF receives activation requests from the AMF for the UP of a PDU session and event messages if necessary.
- Selection of a PCF and signaling with the PCF regarding policies for the PDU session.

However, as shown in Figure 8.15 and according to the statements on the SMF, the direct contact in the 5GC for a UE is the AMF (Access and Mobility Management Function). It also results in the indirect communication of the SMF via AMF. It is also worth mentioning that a PDU session setup is always initiated by the UE, e.g., for a web page request. In case a UE is triggered by the network, e.g., with an incoming call during telephony, there are always-on PDU sessions [88].

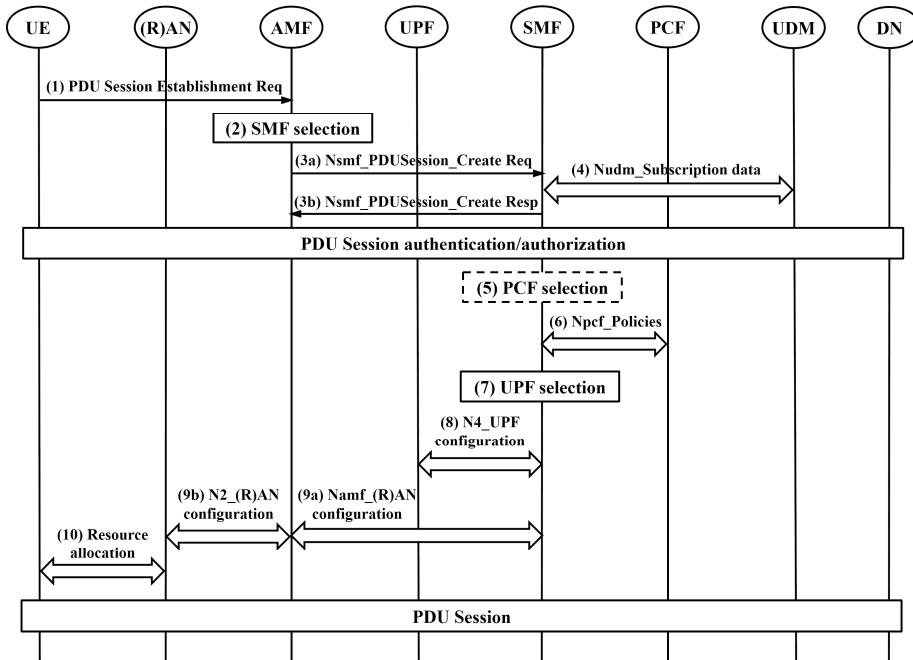


Fig. 8.16: PDU session setup initiated by UE [39]

According to Figure 8.16, the UE sends a (1) PDU Session Establishment Request to the AMF. This selects an SMF in (2) and signals the desired PDU session establishment in (3) via the SBI Nsmf. The SMF in (4) then contacts the UDM (Unified Data Management) via Nudm to retrieve the subscription data. If necessary, the SMF selects a PCF (Policy Control Function) in (5). The SBI Npcf is then used to call up the policies for the desired PDU session in (6). Then the SMF in (7) selects the UPF for this PDU session and allocates an IPv6 address or IPv6 prefix. Next, the UPF is configured via N4 in (8). In (9), the (R)AN is then configured accordingly via the Namf of the AMF and N2. And in (10), the UE is equipped with IPv6 address and prefix and provided with the appropriate QoS rules. The PDU session is established, and user data packets can be transmitted from the UE to the DN (Data Network) and vice versa [39; 88].

The interrelationships in the flow, according to Figure 8.16, are further clarified by the signaling messages from network practice shown below. Figure 8.17 shows a message (3a) captured at the API Nsmf between AMF (in the practical example with the IP address 10.0.1.20) and SMF (IP address 10.0.1.25) to request a PDU session. It contains a JSON document (JavaScript Object Notation) with various information about the user, e.g., the identification specified for their SIM card in the form of the SUPI (Subscription Permanent Identifier, here imsi-00101000000001), the destina-

tion network DNN (here Internet), etc. In addition, this message transports the NAS SM information received and forwarded by the AMF via the N1 interface from the UE. This specifies the desired PDU session in more detail, among other things, concerning uplink and downlink bit rates (maximum possible here), PDU session type (IPv4 here), and SSC mode (Session and Service Continuity) to ensure service continuity (SSC Mode 1 here). Figure 8.18 shows a message (4) with the subscription data sent from the UDM (IP address 10.0.1.22) to the SMF (IP address 10.0.1.25) that was subsequently captured at the API Nudm. Again, the information is delivered in the form of a JSON document. Included are S-NSSAI with the SST (Slice/Service Type) for the network slice (here SST = 1, eMBB application), PDU session type (here IPv4, IPv6), SSC mode (here SSC mode 1), 5QI (5G QoS Identifier) as a reference for handling the QoS flow (here 5QI = 9 for WWW service), ARP (Allocation and Retention Priority) for defining the relative importance of a QoS flow (here ARP = 8, i.e., medium priority), and the session AMBR (Aggregate Maximum Bitrate) concerning the maximum aggregate bit rate of the PDU session (here approx. 1 Gbit/s in the uplink and downlink). Based on these settings, the SMF (IP address 10.0.1.25) programs the UPF (IP address 10.0.1.26) via the N4 interface using the PFCP message (8) shown in Figure 8.19 (see Section 8.1). Concluding this example from network practice, Figure 8.20 shows the message (9a) used by the SMF (IP address 10.0.1.25) to deliver the RAN and UE configuration data to the AMF (IP address 10.0.1.20). The information transmitted in this procedure includes the PDU session type (here IPv4), the session AMBR (here 1 Gbit/s), the SST for the slice (here 1, eMBB), and the PDU address representing the application for the UE (here IP address 10.45.0.2) [37].

```

Internet Protocol Version 4, Src: AMF (10.0.1.20), Dst: SMF (10.0.1.25)
Transmission Control Protocol, Src Port: 59888, Dst Port: 7777, Seq: 216, Ack: 16, Len: 1133
HyperText Transfer Protocol 2
> Stream: DATA, Stream ID: 1, Length 1124
✓ MIME Multipart Media Encapsulation, Type: multipart/related, Boundary: "=FXv8V2Y6XN+j03hVa52Wzg="
    [Type: multipart/related]
    First boundary: ---FXv8V2Y6XN+j03hVa52Wzg==\r\n
    ✓ Encapsulated multipart part: (application/json)
        Content-Type: application/json\r\n\r\n\r\n
    ✓ JavaScript Object Notation: application/json
        ✓ Object
            ✓ Member: supi
                [Path with value: /supi:imsi-00101000000001]
                [Member with value: supi:imsi-00101000000001]
                String value: imsi-00101000000001
                Key: supi
                [Path: /supi]
            > Member: pei
            > Member: pduSessionId
        ✓ Member: dnn
            [Path with value: /dnn:internet]
            [Member with value: dnn:internet]
            String value: internet
            Key: dnn
            [Path: /dnn]
        > Member: sNssai
        > Member: servingNfId
        > Member: guami
        > Member: servingNetwork
        > Member: n1SmMsg
        > Member: anType
        > Member: ratType
        > Member: ueLocation
        > Member: ueTimeZone
        > Member: smContextStatusUri
        > Member: pcfId
    Boundary: \r\n---FXv8V2Y6XN+j03hVa52Wzg==\r\n
    ✓ Encapsulated multipart part: (application/vnd.3gpp.5gnas)
        Content-Id: 5gnas-sm\r\n
        Content-Type: application/vnd.3gpp.5gnas\r\n\r\n\r\n
    ✓ Non-Access-Stratum 5GS (NAS)PDU
        ✓ Plain NAS 5GS Message
            Extended protocol discriminator: 5G session management messages (46)
            PDU session identity: PDU session identity value 1 (1)
            Procedure transaction identity: 1
            Message type: PDU session establishment request (0xc1)
        ✓ Integrity protection maximum data rate
            Integrity protection maximum data rate for uplink: Full data rate (255)
            Integrity protection maximum data rate for downlink: Full data rate (255)
        ✓ PDU session type
            1001 .... = Element ID: 0x9-
            .... .001 = PDU session type: IPv4 (1)
        ✓ SSC mode
            1010 .... = Element ID: 0xa-
            .... .001 = SSC mode: SSC mode 1 (1)
        > 5GSM capability
        > Extended protocol configuration options

```

Fig. 8.17: Message (3a) captured with protocol analysis software at API Nsmf

```

Internet Protocol Version 4, Src: UDM (10.0.1.22), Dst: SMF (10.0.1.25)
Transmission Control Protocol, Src Port: 7777, Dst Port: 42800, Seq: 91, Ack: 201, Len: 635
HyperText Transfer Protocol 2
    > Stream: DATA, Stream ID: 1, Length 626
    > JavaScript Object Notation: application/json
        < Array
            < Object
                < Member: singleNssai
                    < Object
                        < Member: sst
                            [Path with value: /[]/singleNssai/sst:1]
                            [Member with value: sst:1]
                            Number value: 1
                            Key: sst
                            [Path: /[]/singleNssai/sst]
                        Key: singleNssai
                        [Path: /[]/singleNssai]
                    > Member: dnConfigurations
                    < Object
                        < Member: internet
                            < Object
                                < Member: pduSessionTypes
                                    < Object
                                        < Member: defaultSessionType
                                            [Path with value: /[]/dnConfigurations/internet/pduSessionTypes/defaultSessionType:IPV4V6]
                                            [Member with value: defaultSessionType:IPV4V6]
                                            String value: IPV4V6
                                            Key: defaultSessionType
                                            [Path: /[]/dnConfigurations/internet/pduSessionTypes/defaultSessionType]
                                        > Member: allowedSessionTypes
                                        Key: pduSessionTypes
                                        [Path: /[]/dnConfigurations/internet/pduSessionTypes]
                                    > Member: sscModes
                                    < Object
                                        < Member: defaultSscMode
                                            [Path with value: /[]/dnConfigurations/internet/sscModes/defaultSscMode:ssc_MODE_1]
                                            [Member with value: defaultSscMode:ssc_MODE_1]
                                            String value: SSC_MODE_1
                                            Key: defaultSscMode
                                            [Path: /[]/dnConfigurations/internet/sscModes/defaultSscMode]
                                        > Member: allowedSscModes
                                        Key: sscModes
                                        [Path: /[]/dnConfigurations/internet/sscModes]
                                    > Member: 5gQosProfile
                                    < Object
                                        < Member: 5qi
                                            [Path with value: /[]/dnConfigurations/internet/5gQosProfile/5qi:9]
                                            [Member with value: 5qi:9]
                                            Number value: 9
                                            Key: 5qi
                                            [Path: /[]/dnConfigurations/internet/5gQosProfile/5qi]
                                        > Member: arp
                                        < Object
                                            < Member: priorityLevel
                                                [Path with value: /[]/dnConfigurations/internet/5gQosProfile/arp/priorityLevel:8]
                                                [Member with value: priorityLevel:8]
                                                Number value: 8
                                                Key: priorityLevel
                                                [Path: /[]/dnConfigurations/internet/5gQosProfile/arp/priorityLevel]
                                            > Member: preemptCap
                                            > Member: preemptVuln
                                            Key: arp
                                            [Path: /[]/dnConfigurations/internet/5gQosProfile/arp]
                                        > Member: priorityLevel
                                        Key: 5gQosProfile
                                        [Path: /[]/dnConfigurations/internet/5gQosProfile]
                                    > Member: sessionAmbr
                                    < Object
                                        < Member: uplink
                                            [Path with value: /[]/dnConfigurations/internet/sessionAmbr/uplink:1048576 Kbps]
                                            [Member with value: uplink:1048576 Kbps]
                                            String value: 1048576 Kbps
                                            Key: uplink
                                            [Path: /[]/dnConfigurations/internet/sessionAmbr/uplink]
                                        > Member: downlink
                                        Key: sessionAmbr
                                        [Path: /[]/dnConfigurations/internet/sessionAmbr]
                                    Key: internet
                                    [Path: /[]/dnConfigurations/internet]
                                > Member: internet
                            > Member: internet
                        > Member: internet
                    > Member: internet
                > Member: internet
            > Member: internet
        > Member: internet
    > Member: internet

```

Fig. 8.18: Message (4) captured with protocol analysis software at API Nudm

```

Internet Protocol Version 4, Src: SMF (10.0.1.25), Dst: UPF-N4 (10.0.1.26)
User Datagram Protocol, Src Port: 8805, Dst Port: 8805
Packet Forwarding Control Protocol
  > Flags: 0x21, SEID (S)
    Message Type: PFCP Session Establishment Request (50)
    Length: 513
    SEID: 0x0000000000000000
    Sequence Number: 5
    Spare: 0
  > Node ID : IPv4 address: 10.0.1.25
  > F-SEID : SEID: 0x0000000000000001, IPv4 10.0.1.25
  > Create PDR : [Grouped IE]: PDR ID: 1
  < Create PDR : [Grouped IE]: PDR ID: 2
    IE Type: Create PDR (1)
    IE Length: 69
    > PDR ID : 2
    > Precedence : 255
    < PDI : [Grouped IE]
      IE Type: PDI (2)
      IE Length: 29
      > Source Interface : Access
      > F-TEID :
      > Network Instance : internet
      > QFI :
      > Outer Header Removal : GTP-U/UDP/IPv4
      > FAR ID : Dynamic by CP 2
      > QER ID : Dynamic by CP 1
    > Create PDR : [Grouped IE]: PDR ID: 3
    > Create PDR : [Grouped IE]: PDR ID: 4
    > Create FAR : [Grouped IE]: FAR ID: Dynamic by CP 1
    < Create FAR : [Grouped IE]: FAR ID: Dynamic by CP 2
      IE Type: Create FAR (3)
      IE Length: 22
      > FAR ID : Dynamic by CP 2
    < Apply Action :
      IE Type: Apply Action (44)
      IE Length: 1
      0.... .... = DFRT (Duplicate for Redundant Transmission): False
      .0.... .... = IPMD (IP Multicast Deny): False
      ..0.... .... = IPMA (IP Multicast Accept): False
      ...0.... .... = DUPL (Duplicate): False
      .... 0.... = NOCP (Notify the CP function): False
      .... .0... = BUFF (Buffer): False
      .... .1... = FORW (Forward): True
      .... .0... = DROP (Drop): False
    > Forwarding Parameters : [Grouped IE]
    > Create FAR : [Grouped IE]: FAR ID: Dynamic by CP 3
    > Create URR : [Grouped IE]: URR ID: Dynamic by CP 1
    > Create QER : [Grouped IE]: QER ID: Dynamic by CP 1
    > Create BAR : [Grouped IE]: BAR ID: 1
    > PDN Type : IPv4
  
```

Fig. 8.19: PFCP message (8) captured with protocol analysis software at interface N4

```

Internet Protocol Version 4, Src: SMF (10.0.1.25), Dst: AMF (10.0.1.20)
Transmission Control Protocol, Src Port: 50844, Dst Port: 7777, Seq: 215, Ack: 1, Len: 765
HyperText Transfer Protocol 2
> Stream: DATA, Stream ID: 1, Length 756
▽ MIME Multipart Media Encapsulation, Type: multipart/related, Boundary: "=--3fHLDUDH7gPfV25skwoPrw=="
    [Type: multipart/related]
    First boundary: =--3fHLDUDH7gPfV25skwoPrw==\r\n
    > Encapsulated multipart part: (application/json)
    Boundary: \r\n--=3fHLDUDH7gPfV25skwoPrw==\r\n
    ▽ Encapsulated multipart part: (application/vnd.3gpp.5gnas)
        Content-Id: 5gnas-sm\r\n
        Content-Type: application/vnd.3gpp.5gnas\r\n\r\n
    ▽ Non-Access-Stratum 5GS (NAS)PDU
        ▽ Plain NAS 5GS Message
            Extended protocol discriminator: 5G session management messages (46)
            PDU session identity: PDU session identity value 1 (1)
            Procedure transaction identity: 1
            Message type: PDU session establishment accept (0xc2)
            .001 .... = Selected SSC mode: SSC mode 1 (1)
        ▽ PDU session type - Selected PDU session type
            .... .001 = PDU session type: IPv4 (1)
        > QoS rules - Authorized QoS rules
        ▽ Session-AMBR
            Length: 6
            Unit for Session-AMBR for downlink: value is incremented in multiples of 1 Gbps (11)
            Session-AMBR for downlink: 1 Gbps (1)
            Unit for Session-AMBR for uplink: value is incremented in multiples of 1 Gbps (11)
            Session-AMBR for uplink: 1 Gbps (1)
        ▽ PDU address
            Element ID: 0x29
            Length: 5
            .... 0.... = SMF's IPv6 link local address (SI6LLA): Absent
            .... .001 = PDU session type: IPv4 (1)
            PDU address information: UE (10.45.0.2)
        ▽ S-NSSAI
            Element ID: 0x22
            Length: 1
            Slice/service type (SST): eMBB (1)
        > QoS flow descriptions - Authorized
        > Extended protocol configuration options
        > DNN
    Boundary: \r\n--=3fHLDUDH7gPfV25skwoPrw==\r\n
    ▽ Encapsulated multipart part: (application/vnd.3gpp.ngap)
        Content-Id: ngap-sm\r\n
        Content-Type: application/vnd.3gpp.ngap\r\n\r\n
    > NG Application Protocol

```

Fig. 8.20: Message (9a) captured with protocol analysis software at API Namf

The advantages of the described SBA approach, with its modularity, openness, and flexibility, also apply to roaming. Figure 8.21 shows two 5G systems with the relevant NFs, one for the visited network, the VPLMN (Visited Public Land Mobile Network), and one for the home network, the HPLMN (Home PLMN). The shown roaming scenario assumes a local termination in the VPLMN. It means that the UE uses an application (AF) in the DN directly connected to the VPLMN. It also means that NSSF (Network Slice Selection Function), AMF, SMF, and UPF (see PDU Session above), and of course, the AF from the VPLMN are used. Due to the roaming process,

however, the UDM (Unified Data Management) and AUSF (Authentication Server Function) from the HPLMN are responsible for subscriber data and authentication, and the PCF (Policy Control Function) for the UE-specific settings. The NFs vSEPP (visited Security Edge Protection Proxy) and hSEPP (home SEPP) are responsible for secure interaction via N32 [37; 130].

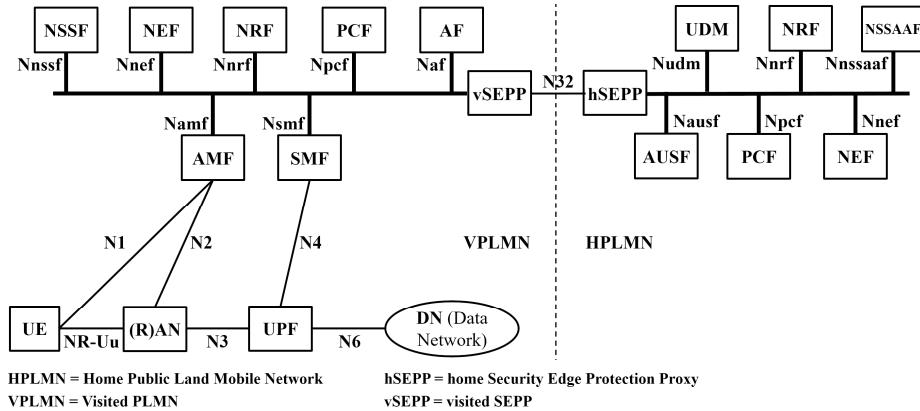
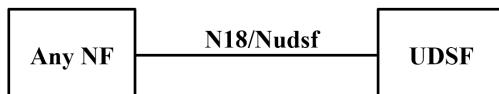


Fig. 8.21: Roaming with local termination on the visited network [37]

Another advantage of the SBA solution chosen for the 5G core is the separation of processing (computing) from data storage. As shown in Figure 8.22, each NF can store its data, e.g., the UE context, in a UDSF (Unstructured Data Storage Function), and retrieve it. A UDSF belongs to the same PLMN as the NF. The data in a UDSF is unstructured and can, therefore, only be interpreted by the corresponding NFs, which increases security. Besides, several NF instances can share one UDSF; the data is then available across NFs. Computing and storage resources are decoupled and can be adapted dynamically; the availability of NFs is correspondingly high [37; 88].



UDSF = Unstructured Data Storage Function

Fig. 8.22: Data storage with UDSF (Unstructured Data Storage Function) [37]

Also, the 5G core provides specific data storage for certain NFs with the UDR (Unified Data Repository), as shown in Figure 8.23. It contains the subscription data of

the users, the policy data for the network, the structured data for the 3rd-party users, and the application-specific data. This repository can be accessed via Nudr from UDM (Unified Data Management), PCF (Policy Control Function), NEF (Network Exposure Function), and indirectly via a NEF from AFs. The data stored here are available in a standardized format. This has the advantage that the UDR can be used internally and externally by NFs from different vendors.

There can be several instances of both the UDR and especially the UDSF. They can be implemented independently of each other or together [37; 88].

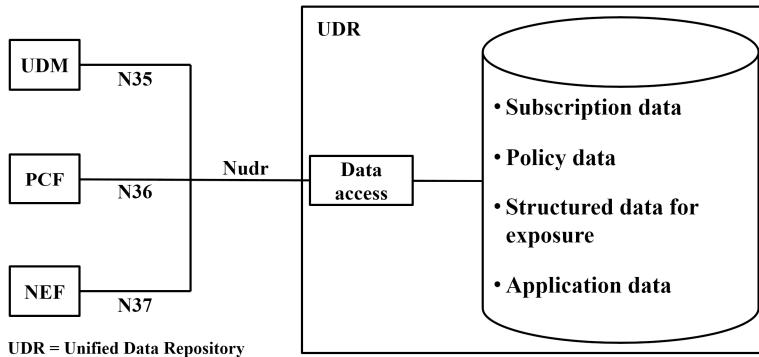


Fig. 8.23: 5GC data storage architecture for SBA [37]

The APIs of the NFs in the SBA, the so-called SBIs (Service-based Interfaces), are specified and implemented as so-called RESTful Interfaces based on the REST architecture (REpresentational State Transfer) [44]. REST is founded on the following principles applied by the 5G SBA [135; 44]:

- A resource, here an NF service profile, can be provided in any form on any server, here NRF, and is represented by a unique URI (Uniform Resource Identifier). Information about a resource is stored in a document, here a JSON document (JavaScript Object Notation), which represents the resource at a certain point in time and can be exchanged between the NFs. Figure 8.18 shows an example from network practice.
- The client-server principle is valid. The client (consumer) requests a service or a resource; the server (producer) provides it.
- Processing in the server is stateless, i.e., the server does not hold any status information. A client request must contain all information necessary for processing. It enables simple load distribution and creates a high degree of resilience.
- Clients can store information received from the server locally in cache memory.

- A client does not know to which server instance it is connected. This is ensured by the NRF.
- The interface concept is uniform. Resources in requests are described by unique addresses, URIs. Not the server, but the specific resource is requested. Therefore, different servers can easily provide the same service. A resource on one server, here, e.g., a service profile on the NRF, can be modified or replaced by a client, here, e.g., an SMF, with a corresponding request. Also, each message exchanged between a client and a server must contain sufficient information on processing the message. The REST architecture also provides that a client can be informed about further possible actions regarding resources via hyperlinks in the replies. However, the application of this REST subprinciple is only intended for later SBA versions.

In the case of the 5G SBA, the RESTful APIs use the protocol stack shown in Figure 8.24, using HTTP/2 (Hypertext Transfer Protocol version 2) as the application protocol, an improved and extended version of HTTP. TCP is the transport protocol whereby secure transmission with TLS is recommended. JSON documents represent the actual applications [43; 135]. Figures 8.17, 8.18, and 8.20 show examples of messages using this RESTful API protocol stack.

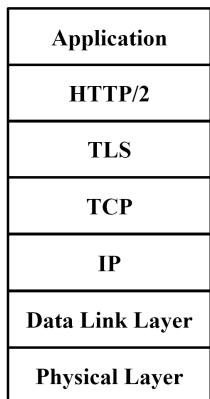


Fig. 8.24: SBI protocol stack [43]

The high degree of modularization provided by the SBA with the NFs communicating via standardized APIs has the additional advantage for a network operator, apart from the advantages mentioned above, that theoretically, each NF can be used from a different manufacturer. In order not to have to perform too many interoperability tests, in practice, no individual NF but bundles of closely related NFs are or-

dered from different manufacturers, e.g., AMF, SMF, and UPF from manufacturer 1, UDM, and UDR from manufacturer 2, and NEF, and NWDAF from manufacturer 3.

As shown in Figure 8.15 and explained in Section 8.2, AFs, i.e., applications provided internally by the 5G network operator itself or externally by 3rd party providers, can also use the NFs of the 5GC. This can be done securely via a NEF (Network Exposure Function). Advantageously, a RESTful API is also used here – as shown in Figure 8.25 [135].

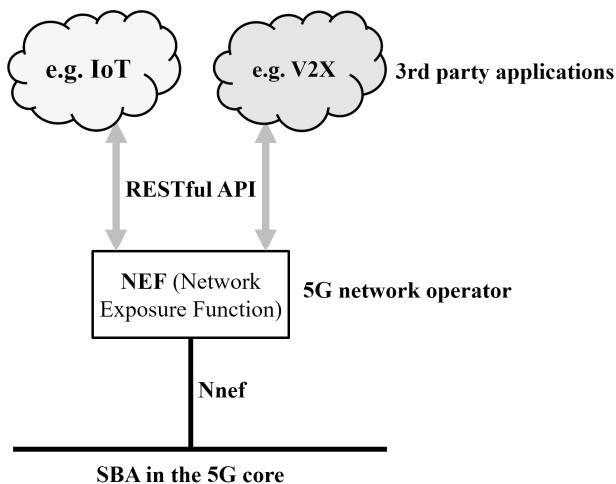


Fig. 8.25: RESTful API for connecting applications to the 5GC via a NEF [135]

With the help of Figure 8.16, the interaction of NFs in the SBA has already been explained in detail above, using the example of the PDU session setup. The understanding achieved through this should now be deepened, considering the knowledge of the RESTful APIs used. Figure 8.26 summarizes three example scenarios, again in a highly simplified manner [135; 39]: a service registration ((1) and (2)), the discovery of a service ((4) and (5)), and again, only in a snippet, the PDU session setup ((3), (6), and (7)).

In the present case, the SMF (Session Management Function) responsible for signaling, and thus for the session context, registers with the NRF (Network Repository Function) and makes itself known and available. It sends a (1) HTTP PUT request as a consumer to the NRF with its service profile in JSON format. Consequently, the NRF creates and stores a URI (<https://...>) to address the SMF's service profile. The HTTP PUT message is answered with a (2) 201 Created Response.

The AMF (Access and Mobility Management Function) acting as the entry point to the 5GC for all CP messages of a UE requires an SMF to handle the signaling and therefore requests a list of available SMFs from the producer NRF via (4) HTTP POST

Request. The AMF informs the NRF in a JSON document about which services the required SMF must support. The NRF searches the registered and stored service profiles and returns the SMFs or URIs matching the request to the AMF in a (5) 200 OK response.

As mentioned above in the context of Figure 8.16, the AMF selects an SMF from the list received and sends a (6) HTTP POST request to it as a consumer to create a session management context for the desired PDU session at the addressed SMF. The required session context is described in a JSON document. If possible, the SMF creates the session management context and confirms this to the AMF with a (7) 201 Created Response. The further procedure for the PDU session setup is then as described above.

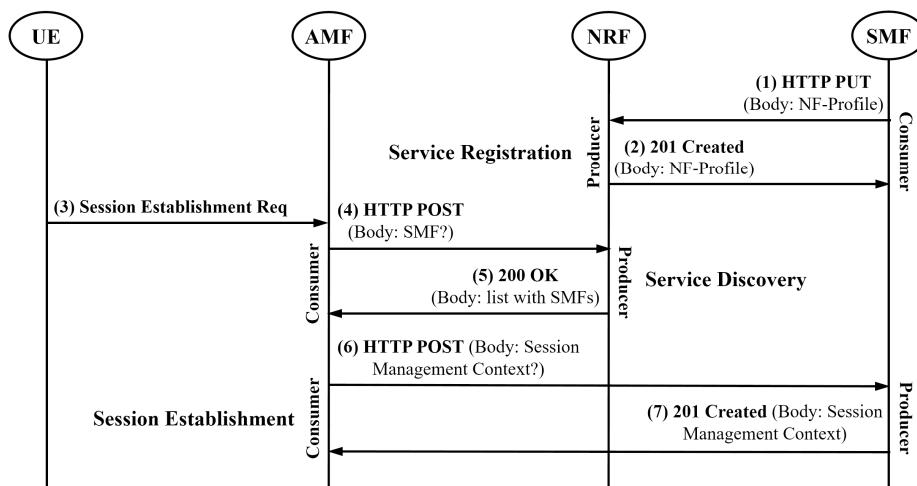


Fig. 8.26: Examples for the interaction of NFs via RESTful APIs in the SBA [135; 39]

8.4 Network Slicing

A fundamental 5G design principle (see Section 6.1) is the Service Based Architecture described in Section 8.3. It provides the basis for comprehensive modularization, which in turn is the prerequisite for being able to compile and combine network functions flexibly as required according to the use cases to be supported (see Chapter 4). In practical implementation, this requires using the design principle of network softwarization, i.e., the application of NFV and SDN for the realization of NF instances and their interaction. If besides, the design principle of multi-tenant capability is to be implemented, network slicing will be useful. Here, two or more logical networks, parallel running network slices, are formed. They enable several tenants, e.g., a mobile network operator, a fixed network operator and an MVNO

(Mobile Virtual Network Operator) for eMBB, a Smart Grid Provider, and a service provider for autonomous vehicles for URLLC, to operate several, here 5, logical communication networks with different characteristics in parallel on one physical network platform.

Figure 8.27 illustrates dependencies. The physical network consists of access networks with specific hardware for transmission technology, hardware routers and switches for networking, as well as computing power and storage resources, primarily based on standard server hardware, preferably in data centers, but also in the access network. This physical network infrastructure, in combination with a virtualization platform, provides an Infrastructure as a Service (IaaS) for the virtual network functions (VNFs) provided and interacting via NFV, SDN, and SBA to form a logical network. In this case, we can talk about NaaS (Network as a Service). If we create several such logical networks on one infrastructure, we get so-called network slices that use the same or different VNFs in specific service chains. Network slicing can extend not only to the core network but also to the access network.

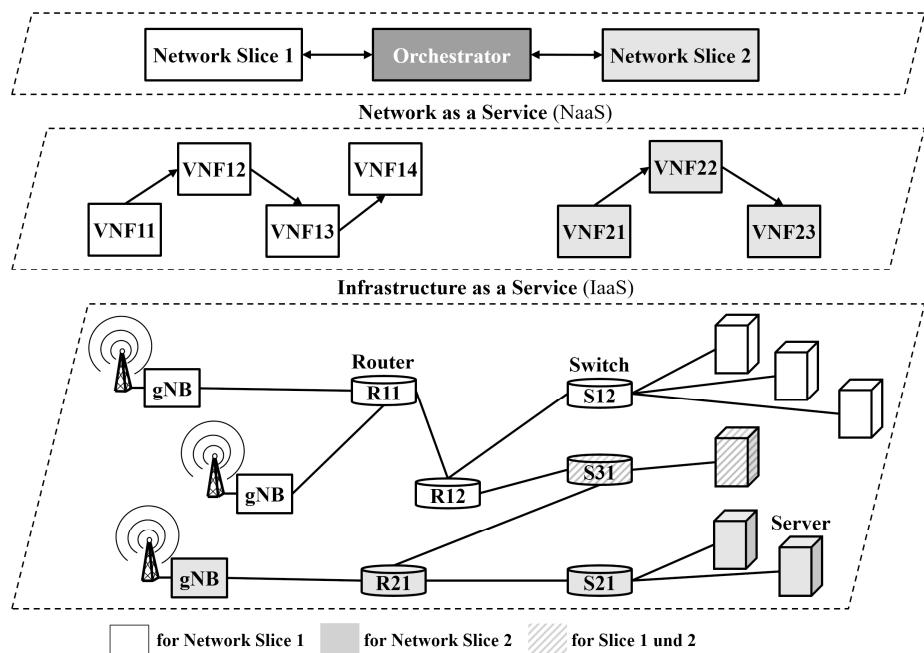


Fig. 8.27: Network slicing based on common physical network infrastructure and virtualization [133]

Figure 8.28 shows the result of network slice generation by orchestrating various CP and UP network functions and RATs for the application cases of smartphones with

high bit rates (eMBB), autonomous driving with low delays and high availability (URLLC), and IoT with very high connection density (mMTC).

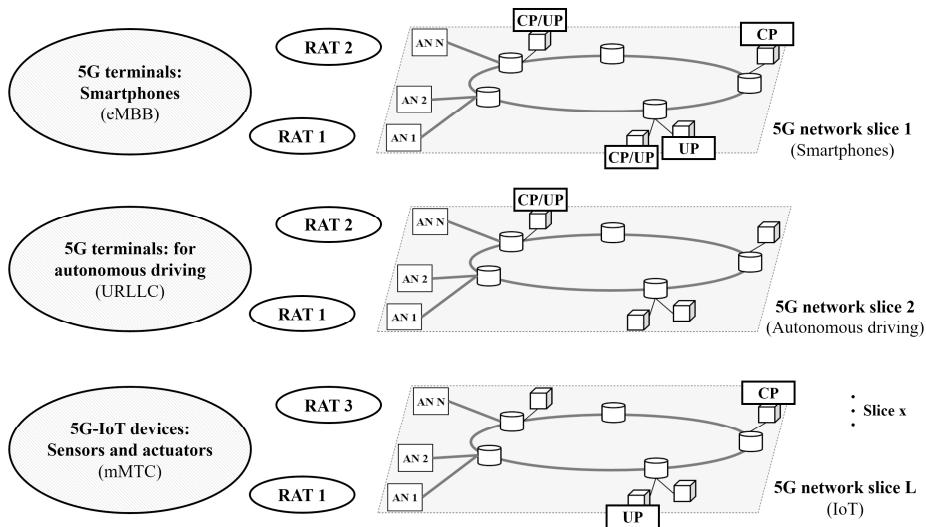


Fig. 8.28: Application scenarios for network slices [140]

Figure 8.29 goes into more detail for a network slice and shows the combination of VNFs and PNFs (Physical Network Function) in VNF Forwarding Graphs or Service Function Chains (see Chapter 3) to form a logical network.

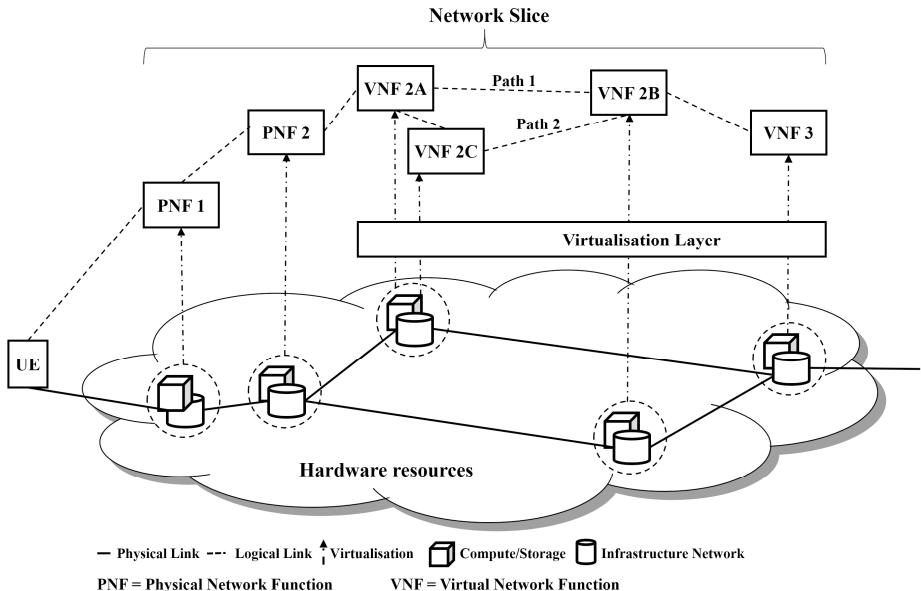


Fig. 8.29: Network slice with orchestrated VNFs and PNFs [140]

Network slicing has the great advantage that tailor-made logical networks based on a single physical infrastructure can be implemented for different applications with a wide range of requirements. These logically separated networks can be assigned to various tenants, i.e., network operators or service providers, and administered and managed by them. In terms of operating costs, however, this is only possible if the networks are orchestrated with a high degree of automation.

Another significant advantage of network slicing is the possibility of considering and handling the resulting different networks in isolation concerning the following aspects [88]:

- Isolation of the network functions: The same or completely different AFs can be used in different slices.
- Isolation of the configurations: Even if the same NFs are used, they can be configured differently in different slices or used in other service chains.
- Isolation of resource usage constraints: In each slice, the specifications for hardware or virtual resources, switch or router throughput, or even latency can be made independently.
- Lifecycle isolation: The generation, modification, and removal of a network slice can be performed independently of any other slice in terms of time. The lifecycles of the NFs within one slice are also entirely independent of those in another slice.
- Isolation of errors: An error that occurs in a network slice is limited to this one.

- Isolation of security areas: Security specifications are individual per slice. Also, an attack on one network slice should not affect another.

According to TS 23.501 [37], an NSI (Network Slice Instance) representing a Network Slice includes CP and UP functions in the 5GC and, if applicable, also in the RAN. A network slice is identified by an S-NSSAI (Single-Network Slice Selection Assistance Information). This comprises an SST (Slice/Service Type) value describing the supported application characteristics according to Table 8.1 and a Slice Differentiator (SD). The latter is optional and only required if multiple slices have the same SST value. A UE can support up to eight slices. Thus, up to eight S-NSAAIs are combined into one NSSAI (Network Slice Selection Assistance Information).

Tab. 8.1: Standardized slice/service types [219]

Slice/service type	SST value	Characteristics
eMBB	1	Slice suitable for enhanced Mobile Broadband services, for high bit rates for mobile use
URLLC	2	Slice suitable for Ultra-Reliable Low Latency Communications services, short delay times, and high reliability
MIoT	3	Slice suitable for Massive Internet of Things applications, for low data volumes per UE, high device density, and extreme coverage
V2X	4	Slice suitable for V2X services, short delay times, high reliability for mobile use, etc.
HMTc (Rel. 17)	5	Slice suitable for High-Performance Machine-Type Communications applications, for short delay times, high availability, and high data rates, but in contrast to V2X, no mobile use

In addition to these specifications, 3GPP proposed in [220] to work with a Generic Slice Template (GST) to describe network slices. With its help, slices with different requirements can be easily described based on standardized attributes. The GST, filled with concrete values for the attributes, becomes the Network Slice Type (NEST), a slice with the desired properties. [220] distinguishes between NESTs standardized by corresponding organizations such as GSMA or 5G-ACIA (S-NEST) and individual private NESTs (P-NEST). The GSMA has taken up this approach and, in [221], has worked out a list of 40 attributes for a GST. Examples of these attributes are availability, delay tolerance, minimum and maximum height of UEs (e.g., for helicopters), throughput in bit/s, energy efficiency, MMTel support (Multimedia Telephony service) with IMS, NB-IoT support, number of UEs, frequency spectrum, QoS, latency or response time between UPF and application server, etc. Based on this, NESTs for the slice/service types from Table 8.1 have already been described in

[221], plus NESTs for “public safety” (Public Safety) and “eMBB with IMS and MPS support” (Multimedia Priority Service). [220] mentions the packet delay budget attribute as an example and lists three slices with packet delays between 1 and 10 ms, 11 and 50 ms, and 51 and 100 ms. Based on such NESTs specified through concrete values for the attributes, the parameters for orchestrating the slices in the 5G core and in the RAN must then be derived.

In the next step, we study the technology of network slicing in a more detailed and practical manner. Please note that an active terminal UE always maintains two types of connections to the 5G core: a signaling connection (NAS) and one or more user data connections (PDU session), in the latter case, with several IP addresses (see Section 7.2). For NAS signaling, there is only one point of contact in the 5GC, an AMF, which also acts as a proxy for participating SMFs. The UE user data is processed in the 5GC by one or more UPFs. However, this also means that when a UE is switched on in the registration phase, an AMF and then the combination(s) of SMF and UPF must be selected (see Section 8.3).

Based on these introductory notes, a network with three exemplary network slices, as shown in Figure 8.30, can now be discussed. On the one hand, there is a relatively simple scenario that UE1 accesses applications on the Internet, referred to as Data Network 1 (DN1), via slice 1. On the other hand, a scenario is shown in which a UE2 accesses two different DNs (DN2 and DN3), providing IoT services using two slices. According to the comments above, UE1 has a NAS signaling connection with AMF1 and a PDU session for the user data with UPF1, controlled by SMF1. In the case of UE2, there is also only one AMF, the AMF2, but for the two PDU sessions, a combination of SMF and UPF, SMF2 and UPF2 for slice 2, and SMF3 and UPF3 for slice 3. In general, a UE can maintain several PDU sessions via several network slices to several DNs, or via one slice into several DNs, or to several slices into one DN. The variants are differentiated by the combination of slice identifier S-NSSAI and DNN (Data Network Name). The latter identifies the target DN [103].

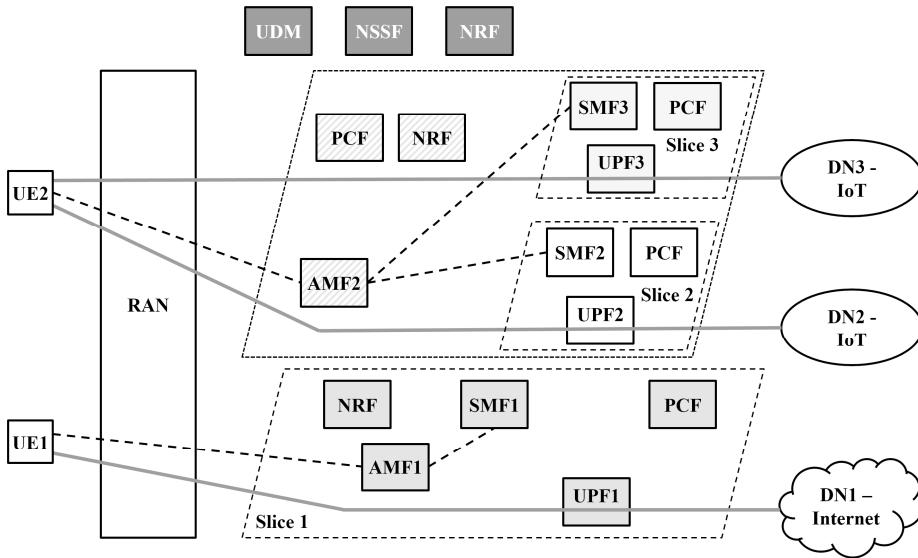


Fig. 8.30: Examples for network slicing in 5G core [103]

The specifications of the network operator for which application, which DN, and which network slice is to be used are transmitted to the UE during the registration process but can also be configured directly in the UE. Changes can also be made using the NAS signaling. The selection of the appropriate network slice according to the operator's specifications is usually done centrally by the NSSF (Network Slice Selection Function) but can also be configured directly in each AMF. As already explained in Sections 8.2 and 8.3, the NRF (Network Repository Function) is used to easily find a required NF, e.g., SMF, UPF, and PCF. As shown in Figure 8.30, the NRF can be made available for specific slices or all slices together. The first approach has the advantage that the isolation between network slices mentioned above as an advantage is complete, and configurations remain invisible across slices. Figure 8.30 does not show the RAN slicing supported by a 5GC system as of 3GPP Release 17 [218]. For this purpose, the NSSAIs corresponding to the PDU sessions are transferred to the RAN. It can then manage the scheduling of IP packets and the allocation of radio resources in the uplink and downlink so that the available resources in the RAN are allocated to the network slices according to the operator's specifications. An orchestration system, NFV-MANO (see Sections 3.1 and 3.2), controls the instantiation, operation, and deletion of a network slice [103].

As mentioned above, UE and the corresponding network slice are linked at the UE registration. Figure 8.31 shows a simplified process for this, assuming that the NSSAI was communicated to the UE by configuration. Accordingly, in the first step, the UE informs the initial AMF via (R)AN (1) about the NSSAI for the requested net-

work slice through (2) Registration Request. The AMF contacts in (3) via Nudm the UDM (Unified Data Management) to retrieve the subscriber data, and with step (4) one or more S-NSSAIs (Single-NSSAI) for this user profile. With this information, the AMF in (5) contacts the NSSF (Network Slice Selection Function) via the SBI Nnssf to select the corresponding Network Slice Instance (NSI). Before the NSSF announces the NSI, the NSSF clarifies in steps (6) and (7) whether the initial or which AMF instance can cooperate with the desired S-NSSAI by contacting the NRF via Nnrf. If necessary, the AMF instance is changed. In (8), the NSSF transfers the selected NSI with permitted S-NSSAI to the appropriate AMF. In turn, it accepts the UE registration for the requested NSSAI and informs the UE of this, including the S-NSSAI, in steps (9) and (10). Subsequently, the UE communicates with and via the selected Network Slice [88; 39].

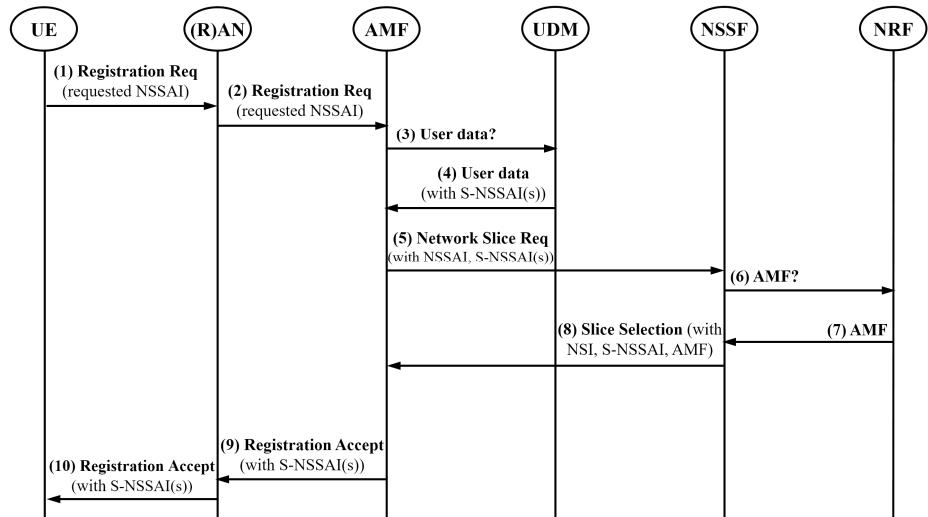
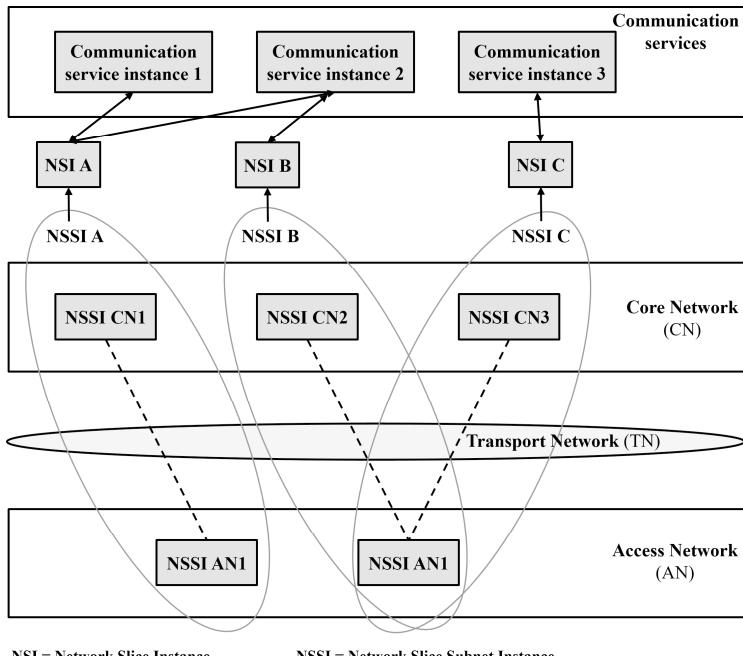


Fig. 8.31: Selecting a network slice [88; 39]

Finally, Figure 8.32 shows how important the 5G design principle of modularity is at the network slice level. According to [42], an NSI (Network Slice Instance) can consist of several NSSIs (Network Slice Subnet Instance), each with its lifecycle, in this case, an NSSI CN of the core network and an NSSI AN of the access network. Together, both subnetwork instances form a network slice instance, including the Transport Network (TN) for interconnection. An NSSI can also be part of two different NSIs. Besides, the same communication service can be provided via different NSIs [42].



NSI = Network Slice Instance

NSSI = Network Slice Subnet Instance

Fig. 8.32: Network Slices formed from Network Slice Subnet Instances (NSSI) [42]

9 5G System

Concerning 5G, RAN has been discussed in more detail in Sections 7.2 and 7.3, and 5G core network in Chapter 8. Due to the new concepts, we highlighted the Service Based Architecture and Network Slicing. This chapter aims to look at a 5G system in its entirety and illustrate from a technical perspective the “great leap forward” for telecommunications networks.

Before diving into this, we will revisit the migration of digital mobile communications networks from the 2nd to the 5th generation (see Section 1.2 and Chapter 2). It also becomes clear that the 5th generation has finally left the field of mobile communications networks and is moving towards a future network with any subscriber access (see Section 3.3).

Figure 9.1 provides an overview of the migration. In the first step, a digital mobile communications network, referred to as 2nd generation, consisted of a circuit-switched core network (CN), the GSM core (Global System for Mobile Communications), and the associated access network (AN). Concerning the easy use of IP over a mobile network, the CN was extended by a packet-switched part, the GPRS core (General Packet Radio Service). In parallel, the AN was migrated to be able to transport IP at medium bit rates with EDGE technology (Enhanced Data Rates for GSM Evolution). This is where the current name GERAN (GSM/EDGE Radio Access Network) originated. The second step was the introduction of a new AN technology, UTRAN (Universal Terrestrial Radio Access Network). In combination with the CN, which continued to consist of the GSM and GPRS core networks, it formed the 3rd generation. It is known as UMTS (Universal Mobile Telecommunications System). Subsequently, the bit rates for UMTS in UTRAN successively increased, and for the first time, mobile terminals were connected via non-3GPP access technology (e.g., WLAN).

The third step resulted in a new, high bit-rate, IP-only access network technology called E-UTRAN (Evolved-UTRAN), mainly known as LTE (Long Term Evolution). It required the provision of a new, real-time IP core network called EPC (Evolved Packet Core) for real-time services such as telephony. This step towards 4G was achieved with 3GPP Release 8. However, the ITU speaks of 4G from IMT-Advanced with LTE-Advanced radio interfaces. With 3GPP, LTE-Advanced is part of Release 10 [54].

During the 3G evolution, the IMS (IP Multimedia Subsystem) was already introduced for Multimedia over IP services in 3GPP Release 5, which then became essential for 4G with LTE and LTE-Advanced because of VoLTE (Voice over LTE), i.e., VoIP in the RAN. It is, therefore, not surprising that the IMS also plays an essential role in the 5th generation of mobile networks, with now really All over IP.

Finally, according to Figure 9.1, 5G provides not only a new, highly modular, and flexible 5G core (5GC) with Service Based Architecture (SBA) and Network Slic-

ing but also a new, extremely powerful RAN technology, NR (New Radio), for very high bit rates, very low latency, and very high connection densities. But that's not all; the 5GC also allows to provide not only NR and non-3GPP WLAN access but also fixed lines via, for example, PON (Passive Optical Network) or DSL (Digital Subscriber Line) and even direct access to a 5G network via a satellite connection. A 5G system can thus really implement FMC (Fixed Mobile Convergence) with only one core network technology. Therefore, as already mentioned above, 5G is actually no longer a mobile network. Instead, if a 5G system is expanded and used in this general way, it can be a new generation convergent network. This consideration was probably the inspiration for the ITU-T's standardization work on future networks (see Section 3.3).

The connecting lines in Figure 9.1 illustrate how the subnetworks operate with one another in different network generations. Today, some network operators have all these subnetworks in parallel. In the future, they will successively shut down the systems of the previous generations. Announcements by mobile network operators indicate that UMTS and UTRAN, and thus the 3rd generation, is phased out first.

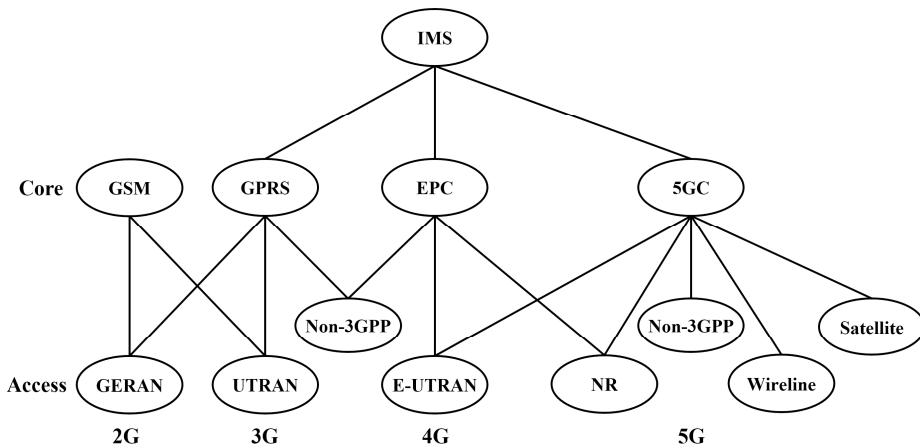


Fig. 9.1: Migration of mobile networks towards 5G

In the following, we will first discuss 4G/5G migration regarding Figure 9.1 (see Section 9.1). Then the already mentioned integration of the IMS in 5G is explained (see section 9.2). A description of the possible FMC follows this with the integration of all access networks (see Section 9.3). These considerations are completed by an overview of IoT support in 5G (see Section 9.4), as well as highlighting the enhanced possibilities offered by the new campus networks added in 5G (see Section 9.5). Finally, this chapter concludes with an overall assessment of a 5G system in a com-

prehensive view of the various concepts and technologies combined in it (see Section 9.6).

9.1 4G/5G Migration

A possible approach to integrate existing 4G technology we have already discussed in Section 7.2. Various options for the introduction of a 5G system were mentioned. With Option 4 in Figure 7.10, an ng-eNB, an LTE base station with an upgraded interface, can be connected to a 5G core via a gNB, i.e., a 5G base station. This allows parallel operation of 4G and 5G RAN technology on a 5GC without any problems. The gNB acts as the master node, and the ng-eNB as the secondary node. In terms of signaling and control, an ng-eNB always communicates via the master gNB using the Xn interface. The user data can also take this route or the direct route to the 5GC via the NG-U interface. Figure 9.2 shows these relationships.

Besides, Figure 9.2 shows the alternative for 4G RAN integration with Option 7, in which a gNB, i.e., a 5G base station as the secondary node, is connected to the 5GC via an ng-eNB, with an LTE base station as the master node. Concerning the interfaces for interconnection, the considerations made above for gNB and ng-eNB are valid.

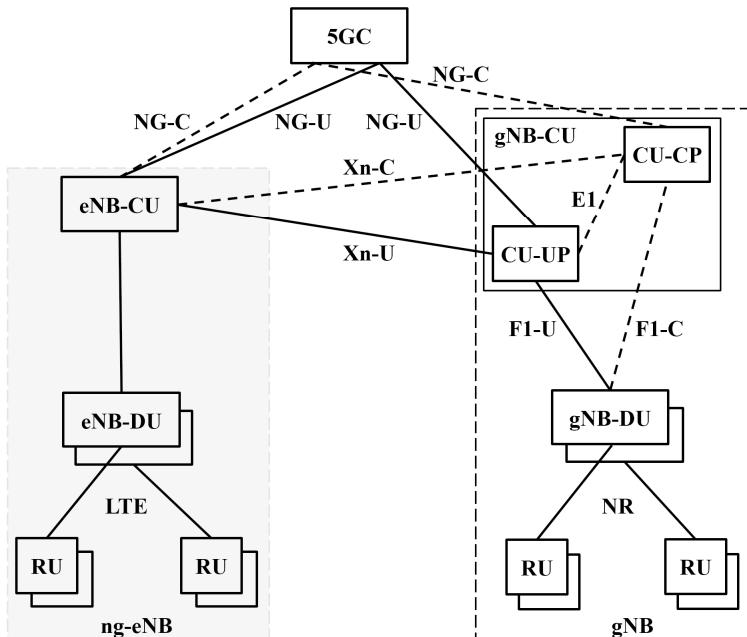


Fig. 9.2: NG-RAN with NR and LTE [129]

In addition, gNBs and ng-eNBs can also be operated independently in parallel on the 5GC. I.e., the 5G RAN can include NR and LTE base stations [129; 19].

Since the integration of 4G into 5G took place at the access network level above, this can also occur in the core network. Please note that in 3GPP Release 14, the CP and UP functions for SGW (Serving Gateway, also S-GW) and PGW (Packet Data Network Gateway, also PDN-GW) in the EPC have been separated and standardized. In the following, it is helpful to consider the correspondences of CN functions in the 4G EPC and 5GC [103]:

- Database for user profiles: HSS (Home Subscriber Server) in EPC – UDM (Unified Data Management) in the 5GC
- Specifications for network behavior (policies), e.g., for QoS: PCRF (Policy and Charging Rules Function) in the EPC – PCF (Policy Control Function) in the 5GC
- User data handling: SGW-U (SGW-User plane) + PGW-U (PGW-User plane) in the EPC – UPF (User Plane Function) in the 5GC
- PDN connection or PDU session control: MME (Mobility Management Entity) + SGW-C (SGW-Control plane) + PGW-C (PGW-Control plane) in the EPC – SMF (Session Management Function) in the 5GC
- Mobility management: MME in the EPC – AMF (Access and Mobility Management Function) in the 5GC.

For 4G/5G integration at the core network level, according to [37] and Figure 9.3, there must be NF modules that provide the CN functionalities HSS + UDM, PCRF + PCF, PGW-C + SMF, and PGW-U + UPF in combination. Also, an interface with reference point N26 was specified for the interaction between MME and AMF, for example, for a handover from a 5G to a 4G radio cell. It should also be noted that this type of 4G/5G interworking via N26 is optional, i.e., not mandatory in the 3GPP standardization [37].

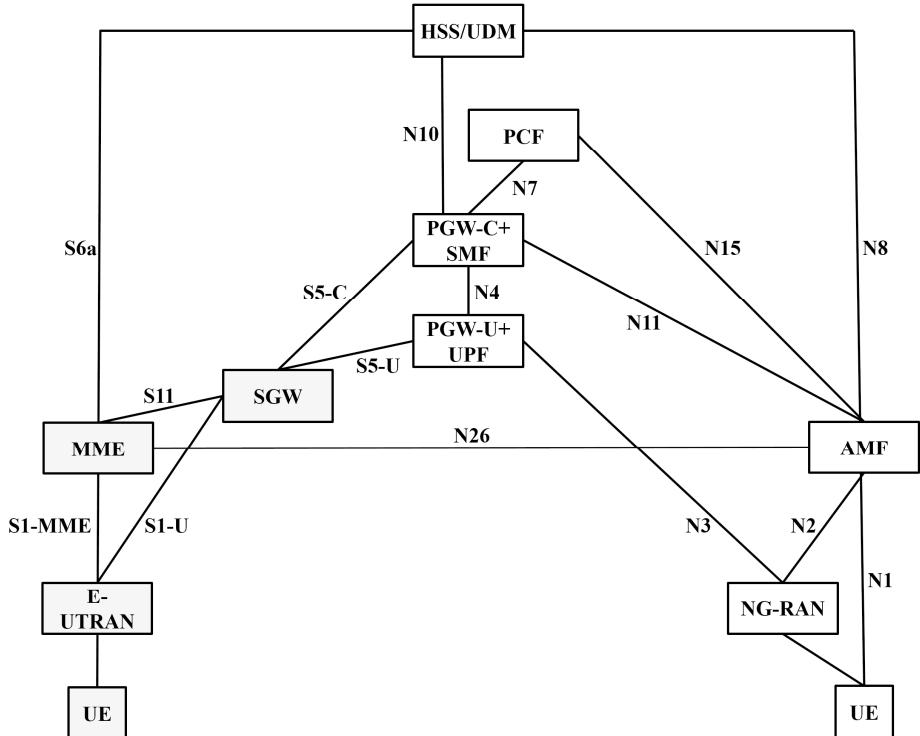


Fig. 9.3: 4G/5G migration through interworking between EPC and 5G core [37]

As already mentioned in Section 7.2 and at the beginning of this section, options for connecting 4G and 5G RANs to 4G and/or 5G core network technology were presented in [40]. These are shown again in Figure 9.4, starting from the pure 4G variant with LTE in E-UTRA connected to an EPC. A next step could then be Option 3 with the addition of 5G NR while retaining the 4G EPC. Alternatively, Option 5 would also allow the switch to the 5G core NGC with further use of the LTE base stations in Evolved E-UTRA. This could be followed by the introduction of 5G-RAN technology NR, according to Option 7 or Option 4. The goal of every migration is, in any case, Option 2 with pure 5G technology.

However, it has already become clear from the comments in Section 7.2 that 5G entry could also be possible via Option 2, a 5G standalone system. A network operator could then have a parallel 4G and 5G solution, which would then be merged using Options 4 or 7.

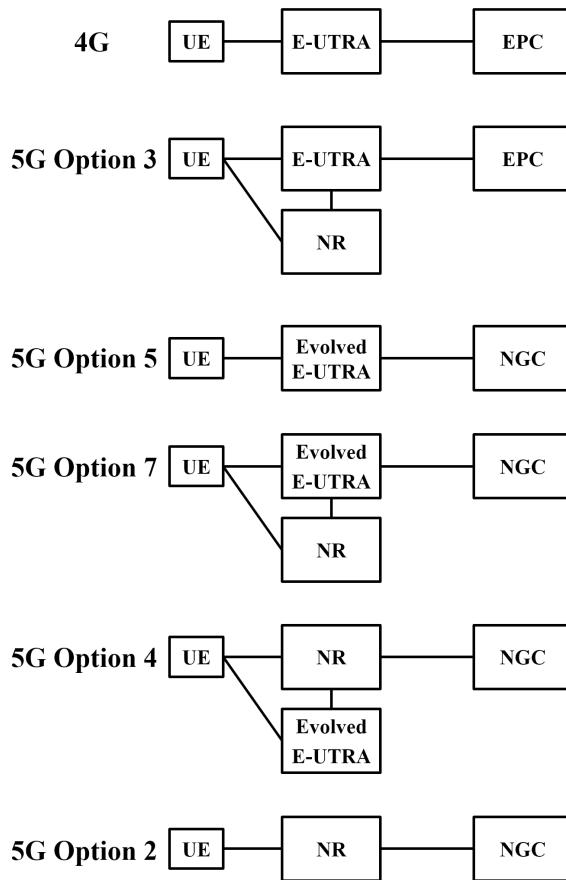


Fig. 9.4: RAN migration from 4G to 5G

Of course, every network operator can choose and follow his own migration path, taking into account his initial situation, his future forecasts, and the optimization of his specific system and operating costs.

9.2 5G and IMS

As mentioned at the beginning of this chapter, the IMS described in more detail in Section 2.2 is the crucial subsystem using SIP signaling to provide VoIP or, more generally, multimedia over IP services in 3G, 4G, and now also 5G mobile networks (see Sections 2.2 and 2.6). With 4G and 5G, it is the only way to support telephony. This, in turn, means that the IMS must be integrated into a 5G system if telephony is offered over 5G.

In the IMS, the HSS (Home Subscriber Server) stores the user profiles and provides the location server functionality required for SIP. In the 5GC, the UDM (Unified Data Management) handles the user profiles or subscription data. Therefore, according to Figure 9.5, IMS integration requires either the collocation of the independently operated HSS with the UDM and their interaction via a direct interface or an integration of the HSS functionality as part of the UDM [33].

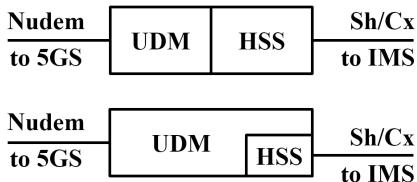


Fig. 9.5: Combination of HSS and UDM for IMS integration [33]

To enable a UE as a SIP user agent to communicate with the IMS for both SIP signaling and RTP real-time user data, IPv6 or IPv4 PDU sessions must be provided via the RAN using the corresponding NFs AMF, SMF, and UPF in the 5GC for the DN (Data Network) of the IMS. Figure 9.6 illustrates this. The resources needed for the required QoS are also reserved. As a result, a UE can exchange signaling messages with the CSCFs in the IMS via SIP, allowing SIP sessions to be established, modified, or terminated. Moreover, user data, e.g., in the form of RTP sessions, can be exchanged with other subscriber terminals via the IMS network elements TrGW/IBCF (Transition Gateway/Interconnection Border Control Function) or a gateway into a circuit-switched network [33].

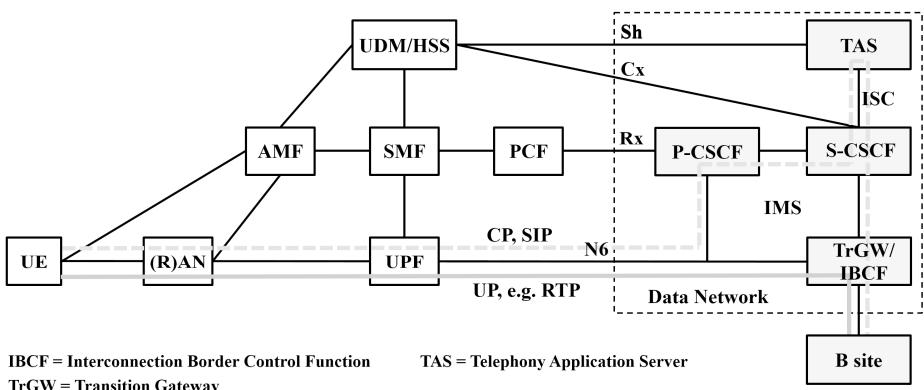


Fig. 9.6: Multimedia over IP with IMS in 5G system [33]

9.3 Access Networks and Fixed Mobile Convergence (FMC)

Chapter 7 was limited to the 5G RAN. The reasons for this were that significant progress had been made in terms of bit rates, latency, and connection densities compared to 4G, but also that Release 15, the first 5G version, concentrated on radio interfaces for 3GPP access with NR technology. Release 16 removed these restrictions and enables the decisive step towards comprehensive Fixed Mobile Convergence (FMC). The fundamental prerequisite for this is applying the 5G design principle “core network decoupled from access network technology”. However, ensuring this is not only the task of the 5GC but also of the access networks. For this reason, various ANs are examined in more detail below concerning their use in a 5G system and the requirements to be met.

In a simplified representation, we can reduce the communication over a 5G system to the used AN, the core NFs AMF, SMF, and UPF in the SBA, and the DN for the application. This simple model is used below to discuss the connection of different AN techniques to the 5GC.

Figure 9.7 illustrates the standard case of 3GPP access with NR or Evolved E-UTRA technology, which has already been mentioned several times, and for which the interfaces and reference points N1, N2, and N3 were planned from the start of the 5G standardization work (see Sections 8.1 and 8.2). As a result, AN and CN harmonize directly here; additional functions for interconnection are not necessary.

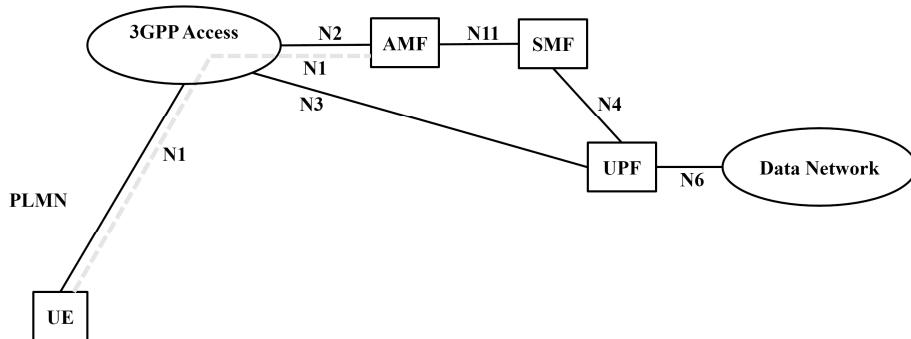


Fig. 9.7: 3GPP access with NR technology [38]

It no longer applies to the two non-3GPP access variants, namely an untrusted AN, e.g., via a WLAN AP on the Internet, or a trusted AN, e.g., via a trusted WLAN AP in a corresponding network environment. Figure 9.8 shows the first-mentioned case with Untrusted Non-3GPP access. A possible application for this would be the VoWifi (Voice over Wifi) with the UE access via an untrusted WLAN AP (e.g., in the Internet) as outlined in Figure 2.32. This does not provide an N2 or N3 interface, and

also no secure access. Therefore an additional NF, the N3IWF (Non-3GPP InterWorking Function), is provided in the 5GC at the interface to the AN, which implements the mentioned reference points and provides an IPsec tunnel endpoint for the UE [38].

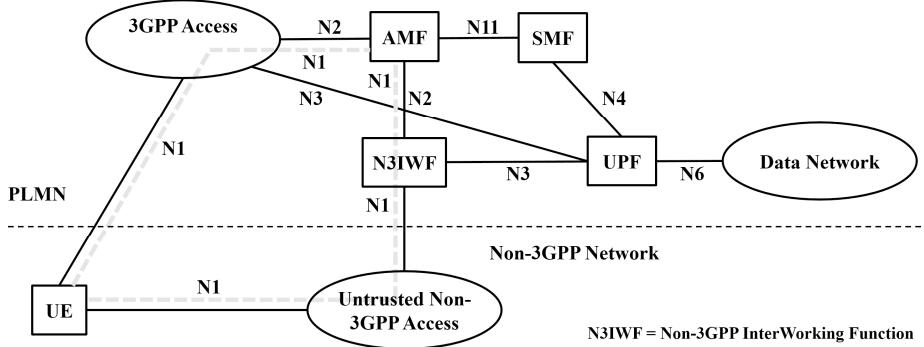


Fig. 9.8: Untrusted Non-3GPP access, e.g., with WLAN [38]

Figure 9.9 describes the case of a Trusted Non-3GPP Access Network (TNAN). Here, too, N2 and N3 interfaces are missing, but a secure connection of the UE is given. In this respect, the additional NF, the TNGF (Trusted Non-3GPP Gateway Function), essentially only has to implement the missing reference points for the 5GC to connect the TNAP (Trusted Non-3GPP Access Point) [38].

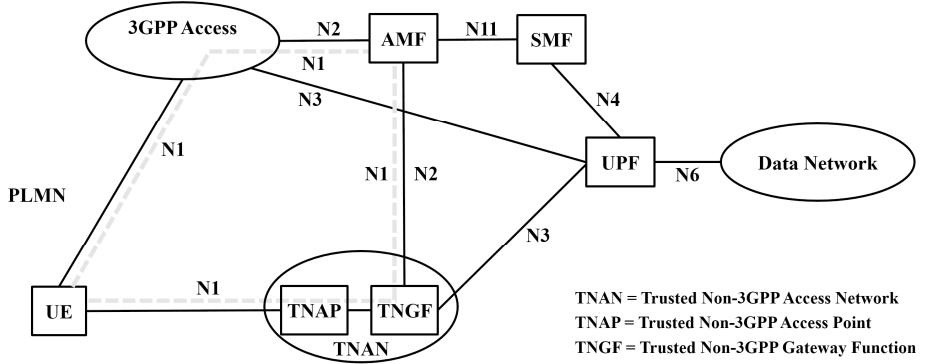


Fig. 9.9: Trusted Non-3GPP access, e.g., with WLAN [38]

With an appropriate offer of untrusted and trusted Non-3GPP Access, a UE or its user decides which AN wants to use. If the decision favors the untrusted AN, the UE first

selects this untrusted AN and then connects to it. Often, several PLMNs (Public Land Mobile Network) can be reached via such AN on the Internet. Then the UE must select the desired 5G network and within this network, an N3IWF. If a trusted AN is to be used instead, the UE first selects the PLMN and then the TNAN. In this scenario, the TNAN used depends on the desired PLMN [38].

Real convergence in a 5G network we will only achieve if the same 5GC can be used to provide users with a wide range of fixed lines, from fixed wireless access (FWA) with mm waves at e.g., 26 GHz, to high bit-rate digital subscriber line (DSL) connections with copper wire pairs, to high bit-rate point-to-point fiber optic interfaces or passive optical networks (PONs).

The Broadband Forum, a global consortium of companies from the telecommunications and IT industries, is responsible for promoting these access network technologies, standardizing them, and taking them into account for 5G [85]. For this reason, the Broadband Forum worked on these issues with 3GPP with the aim of a convergent 5G network. The motivation behind the FMC promotion is as follows [86]:

- AN-independent service experience for customers
- Multi access offers
- Optimized operation of the network
- Uniform technology, training, and services for the operating departments for mobile or fixed subscriber access
- Unified user management
- Use of the core network as far as possible
- Extended range of services via fixed accesses.

On this basis, the Broadband Forum developed six scenarios for FMC in a 5G network, as shown in Figure 9.10. A distinction is made between two types of end systems (Customer Premises Equipment, CPE), here called Residential Gateways (RG): FN-RGs (Fixed Network-RGs) and 5G-RGs. The FN-RGs are already existing systems in the network, such as DSL routers, which do not support 3GPP interfaces and protocols. 5G-RGs, on the other hand, are designed from the beginning for operation on a 5GC.

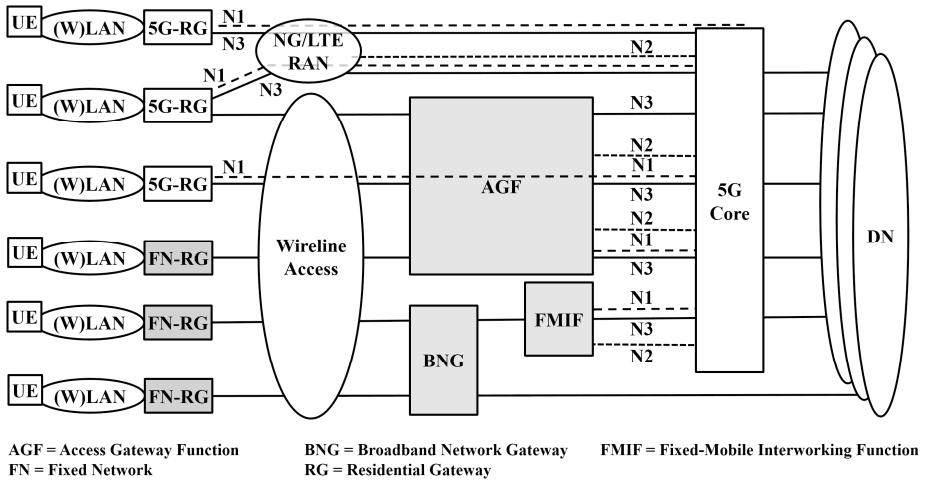


Fig. 9.10: Access network for Fixed Mobile Convergence (FMC) with 5G core [86]

In [86], the six FMC scenarios briefly explained below and outlined in Figure 9.10 are considered:

- Fixed Wireless Access with 5G-RG: The 5G-RGs used here have a 3GPP radio interface and support the 3GPP interfaces N1, N2, and N3. It means that such 5G-RGs can communicate directly with the 5GC via 3GPP Access with 5G- or 4G-RAN.
- Multi Access with 5G-RG: Such a 5G-RG has both a 3GPP wireless interface with N1, N2, and N3 support and a wireline interface with an N1 reference point to the 5GC. Both RG interfaces can be active or standby, for aggregated or split traffic. The functions for the N2 and N3 interfaces cannot be provided via a Wireline AN. They must be made available by an intermediate Access Gateway Function (AGF).
- Integration in Direct Mode with 5G-RG: This 5G-RG has only a wireline interface, no 3GPP radio interface. In this respect, it also only offers the N1 reference point, N2 and N3 are supplemented by the AGF.
- Integration in Adaptive Mode with FN-RG: This scenario is very similar to the direct mode integration with 5G-RG. However, the FN-RG does not support an N1 interface either, so the N1, N2, and N3 functions must be provided by the AGF.
- Interworking with FN-RG: This is an FN-RG with a conventional interface that has to be adapted to the 5G environment via a Broadband Network Gateway (BNG). An additional Fixed-Mobile Interworking Function (FMIF) provides the N1, N2, and N3 functions.

- FN-RG in Coexistence: In this case, only an interface adaptation is made by a BNG, which is directly connected to the DN hosting the application, bypassing the 5GC. A typical application for this is IPTV.

Taking up these considerations, 3GPP also knows 5G-RGs and FN-RGs. RGs are systems at the end customer's premises that enable terminal equipment connected to the RG to use voice, data, video, and video-on-demand services, among others. Typical examples are DSL or cable routers.

Also, for 3GPP, a 5G-RG is an end system that provides N1 signaling and security functions for communication with a 5GC. In [38], 3GPP distinguishes between 5G-RGs again in 5G-BRGs (5G Broadband Residential Gateway), described by the Broadband Forum (BBF), and 5G-CRGs (5G Cable Residential Gateway), as they were specified by CableLabs [87] for Hybrid Fiber Coax networks.

Figure 9.11 shows the 3GPP model for the communication of a 5G-RG via a 5G network. Both variants are considered, the 5G-RG with a 3GPP access interface and a second wireline interface or the 5G-RG with only one wireline interface. In both cases, an intermediate Wireline Access Gateway Function (W-AGF) provides the N2 and N3 reference points to the 5GC in the wireline path. With the help of the W-AGF, the wired AN in Figure 9.11 presents the interfaces required for the 5GC. Therefore one refers to a Wireline 5G Access Network (W-5GAN), again distinguishable in W-5GBAN (Wireline 5G BBF Access Network) and W-5GCAN (Wireline 5G Cable Access Network) [38].

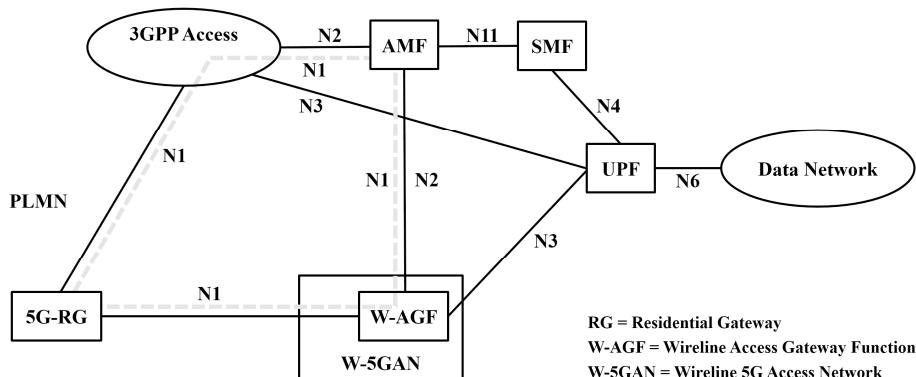


Fig. 9.11: Wireline 5G access network for 5G-RG [38]

Like the BBF above, 3GPP also considers FN-RGs, i.e., end systems with conventional interfaces that do not support N1 signaling for 5G. Therefore, as shown in Figure 9.12, a W-AGF is also used here, which in this scenario implements not only the 5G reference points N2 and N3 but also N1 [38].

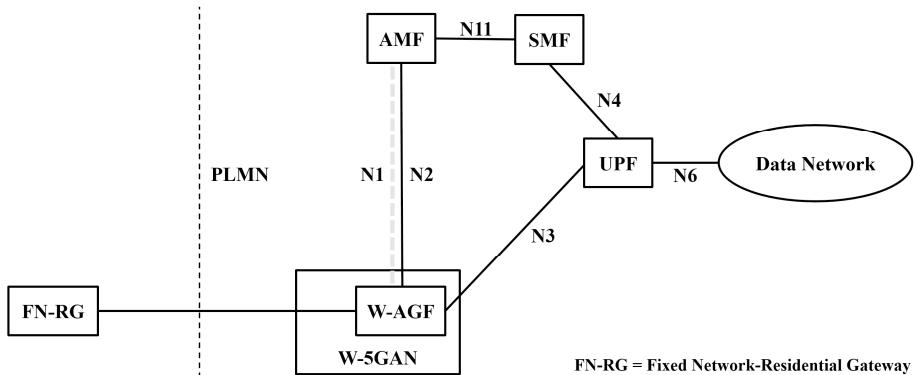


Fig. 9.12: Wireline 5G access network for FN-RG [38]

Besides, 3GPP also knows so-called N5CW end systems (Non-5G-Capable over WLAN) for wireless connections. They do not support the NAS signaling required for direct communication with the 5GC. However, they can still be connected via a Trusted WLAN Access Network and the Trusted WLAN Interworking Function (TWIF) included herein, as shown in [38] and Figure 9.13. During the EAP-based authentication (Extensible Authentication Protocol) of the N5CW UE at the Trusted WLAN Access Point (TWAP), the registration in the PLMN is done simultaneously via TWIF [38].

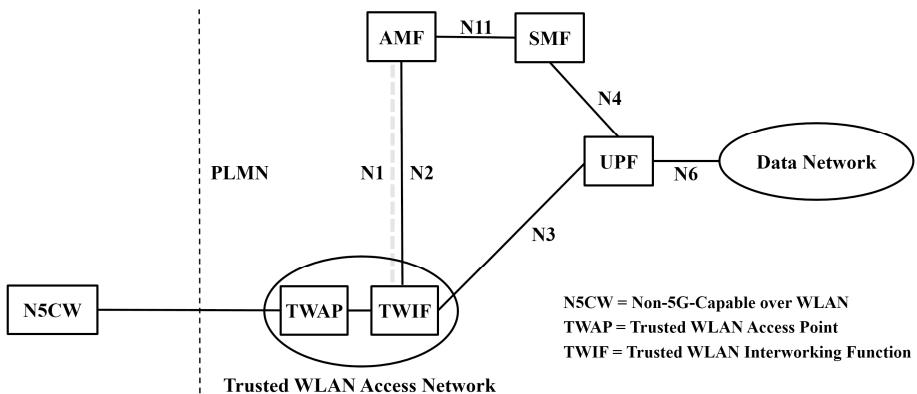


Fig. 9.13: Wireless 5G access network for N5CW-Geräte [38]

To conclude this extensive collection of 3GPP, non-3GPP, and wireline connection options to a 5G network, we would like to mention a highly attractive connection variant introduced with Release 16, the ATSSS (Access Traffic Steering, Switching

and Splitting) function. It allows traffic to be steered, switched, and/or split across various access interfaces. For this purpose, a multi access PDU session is introduced, which can use several parallel interfaces of the different categories – 3GPP, Non-3GPP, and Wireline. For this purpose, the UE must support Multipath TCP (MPTCP) according to Figure 9.14, if necessary (e.g., for Ethernet access), also the ATSSS-LL functionality (ATSSS-Low Layer). The UPF in 5GC must be extended by an MPTCP proxy and, if necessary, by the ATSSS-LL function [38].

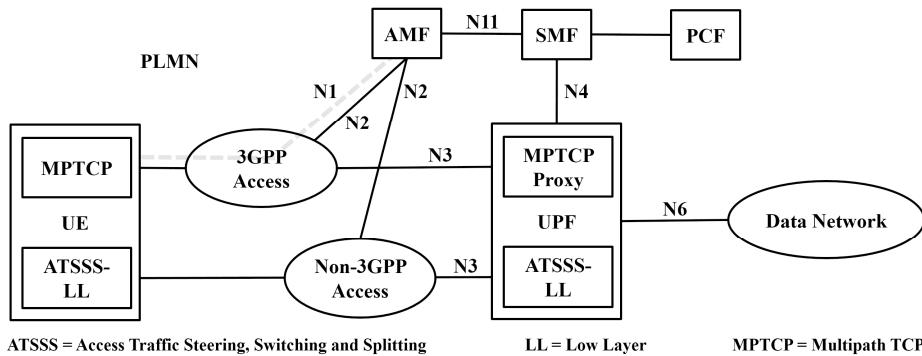


Fig. 9.14: 5G system with ATSSS support [38]

As mentioned above, 5G will include not only terrestrial radio and wireline access interfaces but also satellite-based access with Release 17 to achieve full convergence. The NG-RAN can also be hosted by an Unmanned Aerial System (UAS) or an integrated High Altitude Platform Station (HAPS) instead of a satellite. Carrier units for the latter systems can be uncrewed airships, aircraft, or even drones. Table 9.1 provides an overview of the satellites or platforms that can be used and the distances to be covered [52].

Tab. 9.1: Satellites and UAS for 5G-RAN [52]

Platform	Ground distance [km]
Low-Earth Orbit satellite (LEO)	300 – 1500
Medium-Earth Orbit satellite (MEO)	7000 – 25000
Geostationary Earth Orbit satellite (GEO)	35786
High Elliptical Orbit satellite (HEO)	400 – 50000
UAS incl. HAPS	8 – 50

Figure 9.15 shows the possible connection variants for a 5G-RAN via satellite or UAS.

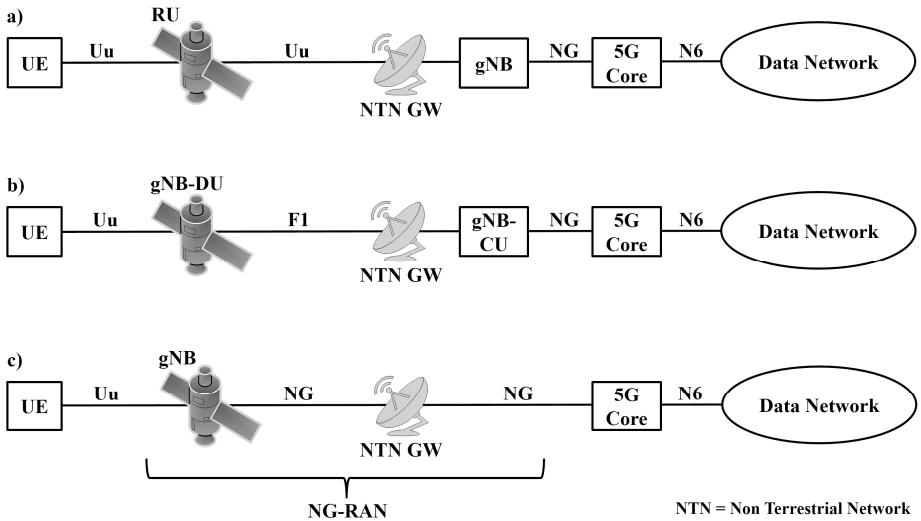


Fig. 9.15: Possible variants for 5G RAN satellite access [52]

In a), only the Radio Unit (RU) is in orbit (see Section 7.2). It represents an analog RF repeater (Radio Frequency) in up- and downlink direction. The gNB on earth provides all other necessary base station functions. The ground station called NTN GW (Non-Terrestrial Network Gateway), and the satellite provide a transparent radio transmission path for 5G. The NR-Uu interface is transparent between the mobile satellite terminal UE and the gNB. In principle, a gNB can also be connected to several satellites.

Connection variant b) in Figure 9.15 uses the standardized possible splitting of the gNB functions into a central CU and a distributed decentralized DU part (see Figure 9.2). The gNB-DU incl. RU is located in the satellite, so the protocol stacks of the lower layers for the transport of user data are also scheduled there. The SRI (Satellite Radio Interface) between NTN GW and the satellite represents the F1 interface in the NG-RAN. Several gNB-DUs onboard different satellites can be connected to one gNB-CU.

Finally, in variant (c), the complete 5G base station, the gNB, is part of the satellite system. Thus, not only the user data but also the signaling is terminated in the satellite. The SRI represents the NG interface (see Section 7.2), the gNB in the satellite implements an interface conversion from NG to NR-Uu [52].

9.4 5G and IoT

As already stated in Chapter 4, an important use case scenario in 5G is mMTC (Massive Machine Type Communications) with low-cost terminals with long battery lifetimes if required, only small amounts of data to be transmitted, but an extreme density of devices in the radio cell [128]. This 5G application area obviously belongs to the IoT (Internet of Things). mMTC, however, only covers a sub-segment of the IoT in the context of 5G. Therefore, the term IoT is first defined before discussing the various IoT segments in a 5G network that go beyond mMTC.

The ITU-T provides an official definition for the Internet of Things (IoT) in Recommendation Y.2060 [222]: „From the perspective of technical standardization, the IoT can be viewed as a global infrastructure for the information society, enabling advanced services by interconnecting (physical and virtual) things based on existing and evolving interoperable information and communication technologies (ICT).“ In addition, [222] notes: „Through the exploitation of identification, data capture, processing and communication capabilities, the IoT makes full use of "things" to offer services to all kinds of applications, while ensuring that security and privacy requirements are fulfilled.“ „Things are objects of the physical world (physical things) or of the information world (virtual world) which are capable of being identified and integrated into communication networks.“

Here it becomes clear that the scope is not limited to the mMTC scenario mentioned in the introduction above and that in the IoT, virtual things are included in addition to physical ones. As for physical networked things, [128] mentions smartphones, sensors, actuators, cameras, vehicles, etc., and points out that they can be low-complexity devices but also highly complex and very powerful devices. The virtual things mentioned in [222] exist in the information world and can therefore be stored, processed, and retrieved. Examples of virtual things are (multimedia) content and application software. Figure 9.16 shows these relationships for a 5G network. In addition, a frequently required IoT gateway is also shown here, enabling the network connection of simple sensors without own communication capabilities.

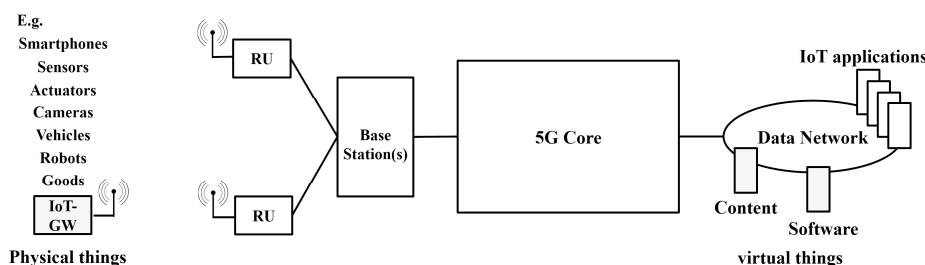


Fig. 9.16: 5G for networking in IoT

As was evident above, MTC and further development mMTC are a subset of IoT. 3GPP now considers this by referring to Cellular IoT instead of MTC in TS 23.501 [37]. In addition, [37] assumes that IoT UEs can be connected via narrowband and broadband radio access networks. For the former, the RAN standards NB-IoT (Narrowband IoT) and LTE-M (LTE for Machines) have already been developed in 4G, and for the latter in 5G NR (New Radio), all for licensed frequency bands.

Starting from this broader view, Ericsson has defined and described four IoT segments in [223] for deployment in 5G:

- Massive IoT
- Broadband IoT
- Critical IoT
- Industrial Automation IoT.

Table 9.2 summarizes their main characteristics.

Massive IoT connectivity addresses large numbers of low-cost, low-bandwidth devices that transmit or receive data infrequently and then only in small amounts. These IoT devices may be challenging to access wirelessly (e.g., electricity meters in a basement), be very numerous (e.g., sensors in a smart city), and may be powered only by batteries (e.g., goods trackers, smart watches). For such mMTC applications, the 3GPP radio technologies NB-IoT and LTE-M have been specified with the characteristics shown in Table 9.2. They enable energy-saving communication, comparatively large transmission ranges, and many simultaneously active devices per radio cell [223].

The broadband IoT application field uses the eMBB capabilities of a 5G system with high data rates and volumes and comparatively low latency for IoT. NR and LTE are used here as radio technologies. This can be network IoT devices with high data rates (e.g., high-resolution surveillance cameras, trains) [223].

Tab. 9.2: IoT segments in 5G and their characteristics [223]

Massive IoT	Broadband IoT	Critical IoT	Industrial Automation IoT
<ul style="list-style-type: none"> - Low-cost devices - Small data volumes - Extreme coverage - Long battery life 	<ul style="list-style-type: none"> - High data rates - Large data volumes - Low latency (for best effort) 	<ul style="list-style-type: none"> - Ultra-reliable data transmission - Ultra-low latency 	<ul style="list-style-type: none"> - Ethernet integration - Time Sensitive Networking (TSN) - Clock synchronization as a service
mMTC	eMBB	URLLC, MEC	URLLC
NB-IoT	NR LTE	NR <ul style="list-style-type: none"> - mmW for high capacity and limited coverage - 1 to 6 GHz for medium coverage and capacity - < 1 GHz for large coverage and limited capacity 	NR LTE
LTE-M		LTE	
<ul style="list-style-type: none"> - 1 Mbit/s UL - 588 kbit/s DL - Support for voice and mobility 			
Meters/counters, sensors, trackers, wearables	Personal cars, commercial vehicles, trains, wearables, gadgets, cameras, sensors, actuators, trackers	AR/VR, autonomous vehicles, mobile robots, real-time human machine collaboration, cloud robotics, haptic feedback, real-time fault prevention, coordination and control of machines and processes	Ethernet with TSN support for various QoS requirements

Critical IoT applications in 5G systems are based on the URLLC use case scenario, where highly reliable data transmission is possible with very short delay times. In contrast to broadband IoT, which achieves low latency on a best-effort basis, data can be delivered within certain latency limits with the required guarantees, even in heavily loaded networks (e.g., for AR/VR, autonomous driving, and collaborative robots). The very low delay times can be achieved by using short slots in the transmission frames on the NR radio link (see Section 7.1), operating the responsible UPF next to the base station (see Section 8.2), and/or using MEC for IoT data handling (see Section 3.1). Highly reliable data transmission can be achieved by redundant transmission paths in the 5G system with redundant UPFs (see Section 8.2) [223].

IoT in industrial applications also uses the functionalities of the URLLC scenario. They enable the integration of wirelessly connected and mobile IoT devices into an industrial company's previously wired communication infrastructure. Based on the URLLC characteristics mentioned above, deterministic real-time communication with the required maximum delay times is ensured despite radio communication integrated into Ethernet networks specially designed for industrial automation. As of 3GPP Release 16, a virtual LAN based on a 5G system can be set up, in which UEs can communicate with each other in VLANs via uni-, multi- or broadcast operation. If the 5G system is part of a time-sensitive networking (TSN) network with real-time requirements for the connected Ethernet networks, it acts as a TSN bridge, as shown in Figure 9.17. For this purpose, the UPF must act as a bridge and provide a Network-Side TSN Translator (NW-TT). A Device-Side TSN Translator (DS-TT) must be connected behind the UE on the terminal side. UE, gNB, UPF, and the TTs are synchronized to the 5G clock. The TTs then ensure that all terminals in a TSN domain, including those connected via 5G, are synchronized based on the TSN clock using the gPTP (generalized Precision Time Protocol) and the time stamps transmitted with it. Accuracies of less than 900 ns can be achieved [37; 224; 223].

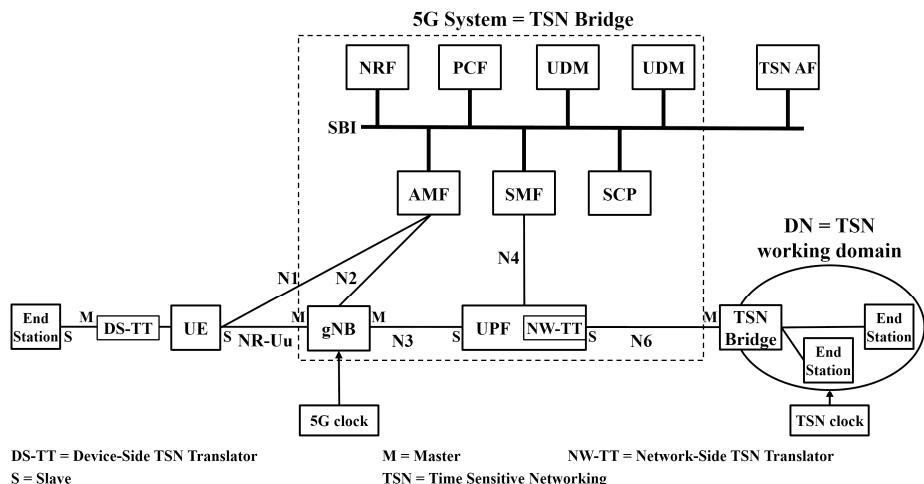


Fig. 9.17: 5G system as TSN bridge for real-time Ethernet [37]

As illustrated in Table 9.2, different radio technologies are used for IoT in 5G according to Figure 9.18: NB-IoT, LTE-M, LTE, and NR. These can operate efficiently side by side because Dynamic Spectrum Sharing is supported, i.e., the frequency spectra originally intended for LTE can be dynamically allocated by NR and the LTE radio technologies, allowing very flexible and efficient use of the total available spectrum

[223]. Figure 9.18 shows the different RATs, the use of MEC, and the ability to deploy IoT- and company-specific IoT network slices.

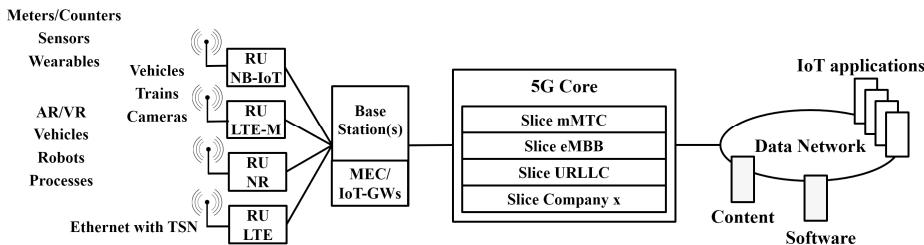


Fig. 9.18: Functionalities of a 5G system for IoT

9.5 5G Campus Networks

Different industries (verticals) with specific applications have different requirements for the communications network and, thus, for a 5G system. One aspect is network coverage. In many cases, wide-area, e.g., nationwide or international network coverage, is required, e.g., for logistics applications. This requirement can typically only be met with a public mobile communications network available over a wide area. For other applications, e.g., for networking a factory site, local coverage is sufficient. In this case, there may be special requirements for data sovereignty, availability, or latency. Specifically for this purpose, non-public 5G networks (NPN, Non-Public Network) with only local coverage were specified by 3GPP in the context of 5G [22; 37], in contrast to public 5G networks (PLMN, Public Land Mobile Network) with, e.g., nationwide coverage [226].

In Germany, the Federal Network Agency (Bundesnetzagentur) has reserved a separate frequency range, 3.7 - 3.8 GHz, for such so-called campus networks. This means that up to 100 MHz of bandwidth can be used exclusively by a company or organization in a defined geographical area after allocation by the regulatory authority. As areas of application for such campus networks, [22] mentions production/industrial automation, logistics in ports and on railroad stations, smart electric power supply, mining, and mobile, i.e., transportable 5G networks in agriculture, on construction sites, and at events.

According to [37], such an NPN can be implemented as a standalone non-public network (SNPN) or as a subnetwork of a public PLMN. In addition, NPN and PLMN can share a common RAN. Not supported by an SNPN are emergency calls, roaming, and handover between different NPNs or NPN and PLMN [37]. Depending on the operator agreements and regulatory requirements, subscribed services in the PLMN can, in principle, be accessed from an NPN and vice versa [22].

Based on the variants outlined in [37] and mentioned above, the following section examines the implementation options for a 5G campus network.

According to Figure 9.19 a), the first variant is an independent private 5G network, an NPN, with network components for base stations (gNB), the 5G core, and possible MEC hosts for edge computing that are entirely independent of the public network, the PLMN. In addition only locally used frequencies in the 3.7 – 3.8 GHz range, license-free frequency ranges (e.g. WLAN), possibly also millimeter waves (mmW) in the 24.25 - 27.5 GHz range are used here for radio transmission. Millimeter waves can only cover comparatively short distances of up to approx. 200 meters. However, this is usually unproblematic for a campus network. Advantageously, more than 1 GHz bandwidth is available in this case. This results in the highest data rates and/or very low latency times for applications. Such a completely independent campus network can be set up and/or operated by the organization, company, or an appropriate service provider.

Variant two, b) in Figure 9.19, implements the NPN as a network slice (see Section 8.4) and thus as a virtual subnet as part of a public PLMN. Although the slice in the 5G core and possibly also in the RAN is used exclusively for internal company communication, the physical network components of the PLMN are used, and the UEs used in the NPN must use SIM cards for the public PLMN and be registered there. In addition, only the publicly allocated frequency spectra, e.g., 3.4 – 3.7 GHz, can be used. However, a clear advantage in this scenario is that simple IoT devices, or IoT devices that are difficult to access due to the attenuation values at higher frequencies, can be connected with the NB-IoT and/or LTE-M radio technology only permitted in PLMNs. A possible step towards an even more customized solution could be to operate UPF and/or (private) MEC on-site in the local network to ensure data sovereignty for the slice. This would be a simple hybrid solution, i.e., a combination of PLMN and NPN components.

Variant three, c) in Figure 9.19, is a hybrid approach with gNBs shared between NPN and PLMN, a so-called shared RAN. This means that a public mobile network operator is responsible for the base stations and could use public licensed frequencies, including NB-IoT and LTE-M, or alternatively provide radio cells specifically for the campus, which would then operate exclusively in the frequency range to be used locally. All other network components, 5G core, MEC host, are in the responsibility of the campus network operator.

The last and fourth variant, d) in Figure 9.19, is also a hybrid solution with independent gNBs for the NPN, connected to the 5G core of a PLMN. Compared to the hybrid solution c), one advantage is the higher flexibility regarding the gNBs for the NPN [225].

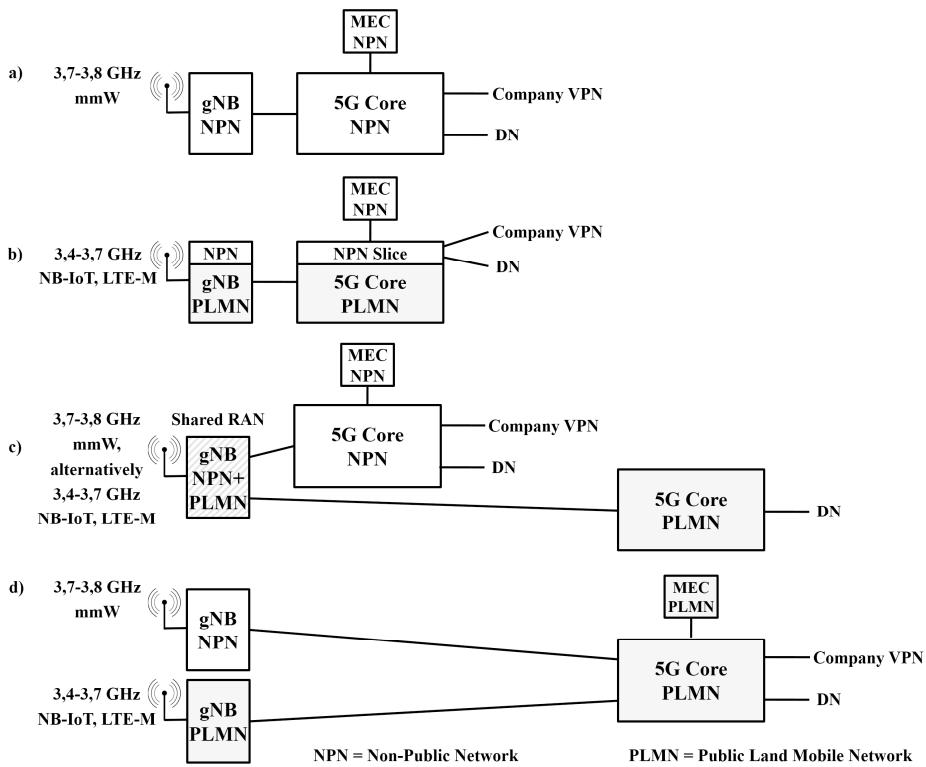


Fig. 9.19: Implementation options for 5G campus networks [225]

Table 9.3 summarizes and compares the four options for implementing a 5G campus network. On the one hand, the essential characteristics are discussed. On the other hand, they are compared based on deployment costs, data sovereignty, security, and flexibility. Although a standalone network generates comparatively high costs, it offers only advantages – except for the impossibility of using NB-IoT and LTE-M. In particular, sovereignty over all exchanged and stored data can argue for such a standalone NPN, e.g., for a 5G campus network at a factory site with sensitive production data [225].

Tab. 9.3: 5G campus network deployments in comparison [225]

	Independent network	Network slice	Hybrid with shared gNBs	Hybrid with separate gNBs
Characteristics	<ul style="list-style-type: none"> - NPN: - gNB - Campus frequencies - 5GC - MEC Host 	<ul style="list-style-type: none"> - NPN: - MEC Host - PLMN: - Slice für gNB und 5GC - Public frequencies - NB-IoT und LTE-M 	<ul style="list-style-type: none"> - Possibly campus frequencies - 5GC - MEC Host - PLMN: - gNB - Possibly public frequencies - NB-IoT and LTE-M 	<ul style="list-style-type: none"> - NPN: - gNB - Campus frequencies - PLMN: - 5GC - MEC Host
Build-up costs	High	Low	Medium	Medium
Data sovereignty	Very high	Low	High	Low
Security	High	High	High	High
Flexibility	High	Low	Medium	Medium

9.6 5G System in an Overall View

In summary, for this chapter, the following is an overall view of a 5G system. Based on the explanations of the various supported subscriber interfaces from Section 9.3 and the resulting convergence, the FMC, the comprehensive view begins with a look at the physical layer, i.e., the network architecture from a hardware perspective.

The 5G network is structured for this purpose into the 5G core, the transport network, the various radio-based, wired, and satellite-based access networks, as well as private or corporate networks. The latter may even be independent 5G networks (see Sections 4.2 and 9.5) with their own frequencies, e.g., in Germany in the 3.7 to 3.8 GHz range (see Section 5.3).

According to Figure 9.20, the hardware in the core network consists of SDN switches and/or high-performance routers with dedicated HW, the servers for the 5GC, and the management and orchestration (MANO) functions. These servers are typically located in data centers and operate with high availability in the central cloud. The transport network based on SDN switches and/or routers connects the data centers with edge cloud locations and access networks. The edge cloud hardware can be used for MEC applications on the one hand, and the cloud-based provision of CU functions in a C-RAN on the other (see Section 3.1). The AN contains a wide variety of access network technologies already described in Section 9.3. Possible arrangements of the NG-RAN components are described in Section 7.2, where the

remotely operated base stations are connected via backhaul transport connections, usually by fiber optics. PONs (Passive Optical Network) are particularly cost-effective for this purpose. Fronthaul connections with optical fiber ensure the interconnection of the RUs (including the antennas) located near the users with the more central functions of the base stations and/or the RAN-DUs with the associated RAN-CUs.

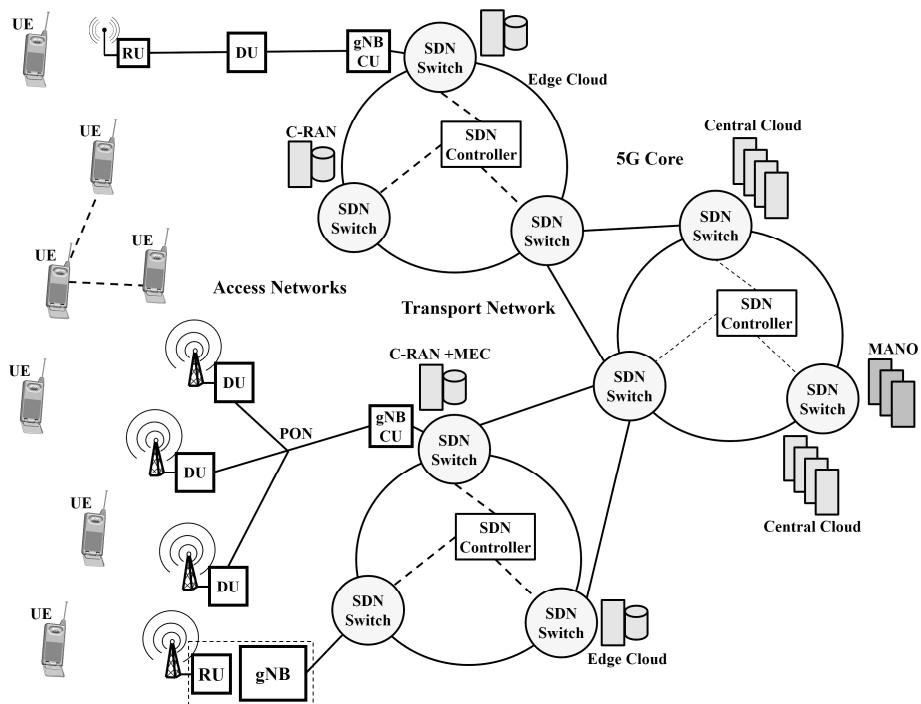


Fig. 9.20: 5G overall network view from the perspective of the physical layer [126]

This L1 architecture, as described above, is the basis for the functions of all higher layers in a 5G system, as shown in Figure 9.21. It starts in the physical layer with the transmission technology HW of the different ANs and the server HW for computing and storage, as well as the HW of the SDN switches and possibly a specific HW for SDN controllers, which in this case would work as PNF (Physical Network Function).

However, the aim is to provide and use the required network functions as VNFs (Virtual Network Functions) wherever possible. To this extent, the HW must be decoupled from the NF software by a virtualization layer, e.g., using a hypervisor. Based on this, virtual computing, storage, and network resources are made available to the VNFs in the form of virtual machines (VM) and/or containers. This could

also include the virtual resources for SDN controllers. Besides, the MEC platforms for the subscriber-oriented hosting of applications are also located at this level.

These virtual resources, VMs, and/or containers enable the instantiation of 5GC functions such as AMF, SMF, UPF, VNFs for SDN controllers, and MEC applications. The SBA works in this layer (see Section 8.3) and supports modularity, flexibility, and building network slices (see Section 8.4). The IMS, as a complex application within a data network (see Section 9.2), is also positioned here. Also, each VNF in this layer has an associated element management function.

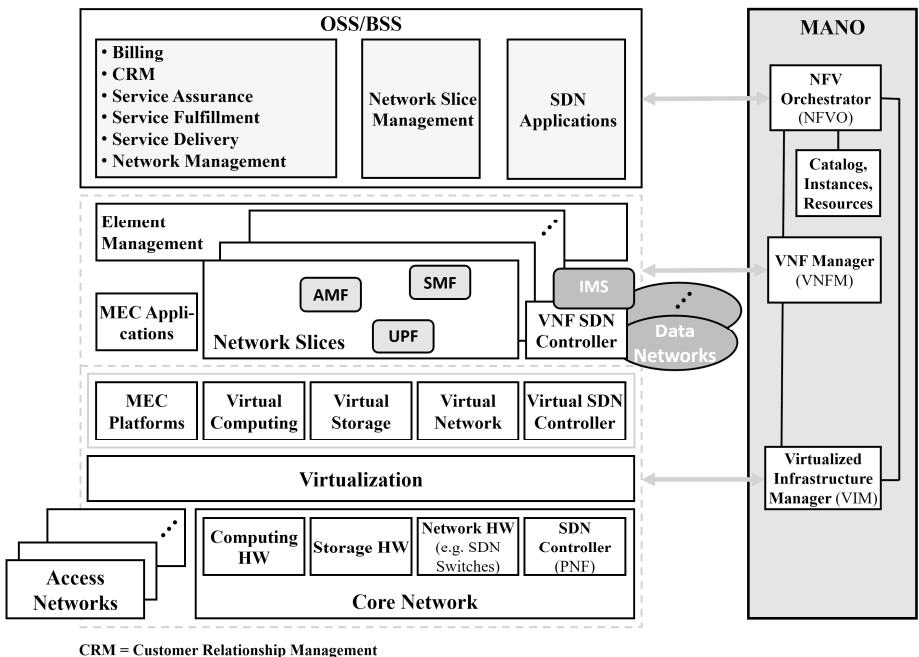


Fig. 9.21: 5G system in an overall view

Based on these, the OSS/BSS (Operations Support System/Business Support System) is used for network management, network slice management, configuration of the SDN controllers via an API for the required SDN applications, provision of communication services, initiation of necessary diagnoses, CRM (Customer Relationship Management) and billing of services [126].

Such a complex system can only be operated with a high degree of automation. The NFV-MANO (NFV-Management and Orchestration) provides the necessary functions. The NFV Orchestrator (NFVO) is responsible for installing and configuring new network services and composing network services out of VNFs. It receives the required information from OSS/BSS and, above all, databases, which provide a

catalog of all NFs and information on the available instances and virtual and physical resources. The VNF Manager (VNFM) also uses these data to manage the lifecycle of a VNF, i.e., instantiation (creating a VNF), update/upgrade (new SW or changed configuration), required scaling (increasing or decreasing the capacity of a VNF, e.g., number of CPUs or VMs) and terminating (returning NFVI resources allocated by a VNF). Finally, the Virtualized Infrastructure Manager (VIM) is responsible for allocating and managing the virtual and physical resources, taking into account the interactions of a VNF with virtual computing, storage, and network resources. Performance, error, and capacity planning data are also recorded (see Section 3.1).

To complement the overall view of a 5G system presented here, it should be noted that in practice, a typical full-service network provider operates a GSM core, a GPRS core, and an EPC including GERAN, UTRAN, and E-UTRA access network technology in parallel to the new 5G system shown in Figures 9.20 and 9.21 (see Chapter 9, Section 2.5 and Figure 2.29).

In conclusion to this 5G overall system overview, the innovations and advantages of 5G compared to 4G and modern wired networks are summarized below. 5G enables a network operator to maintain only one network for all applications and various access networks. In addition to the scenario eMBB with high bit rates, 5G offers an entirely new URLLC with high reliability and low latency, and with mMTC, the connection of a significantly higher number of M2M and IoT devices. In the 5G core network, Service Based Architecture and network slicing introduced real innovations. These were made possible by the consistent use of network softwarization and cloudification. For the first time, new frequency ranges have been opened up for the RAN, on the one hand, between 3.4 and 3.8 GHz, and on the other hand, the millimeter wave range from 24.25 GHz. This requires the use of massive MIMO and beamforming. Another innovative feature is the significantly expanded support for third-party providers with MEC and, above all, the network exposure functions with access to network-internal functionalities. Last but not least, 5G offers companies and organizations the opportunity to set up and operate their campus networks for the first time.

10 5G and Security

For such an ambitious system or network, it is evident that it must be secure. This includes the need to protect the privacy of the users, to ensure the confidentiality and integrity of the messages and data transmitted, and to prevent any kind of cyber attack that could affect the availability and integrity of the network or the confidentiality of the data stored on it. However, this poses a much higher challenge compared to 4G. In addition to the use cases for eMBB (Enhanced Mobile Broadband) with high bit rates, which 4G already supported, there are also use cases for URLLC (Ultra-Reliable and Low Latency Communications) and mMTC (massive Machine Type Communications). This means that in the case of URLLC, successful attacks can cost lives, for example, in 5G-supported autonomous driving or V2X in general. Or that system-relevant infrastructures, such as parts of the power supply network, can fail. On the other hand, mMTC use cases often involve IoT terminals, possibly in large numbers, with a low energy budget due to battery supply and can transfer only small amounts of data. Nevertheless, we have to guarantee security at a high level with appropriate security functions and protocols [88; 159].

Figure 10.1 shows other specific security challenges for 5G. The overall system view presented there was derived from Figure 9.21. It shows that a wide variety of system components, subsystems, infrastructures, platforms, and functions must be considered in the requirements and considerations for security. These include:

- Transport infrastructure with dedicated physical transmission systems
- Wireless and wireline access networks
- Physical and virtual computing resources for the core network, if necessary, also for the access network, e.g., in case of a C-RAN
- NFV with VNFs in a network slice
- Network slices
- MEC platforms and applications
- Management
- Orchestration
- Terminal equipment, the UEs.

This shows that not only the usual security functions for a telecommunications network need to be provided for a 5G system but also that special attention needs to be paid to virtualization in this context because NFV, including C-RAN and MEC, are implemented in cloud environments. Particular attention should therefore be paid to security in the central and edge-cloud infrastructures. The underlying transport network uses SDN, so in particular, the SDN controllers must be secured.

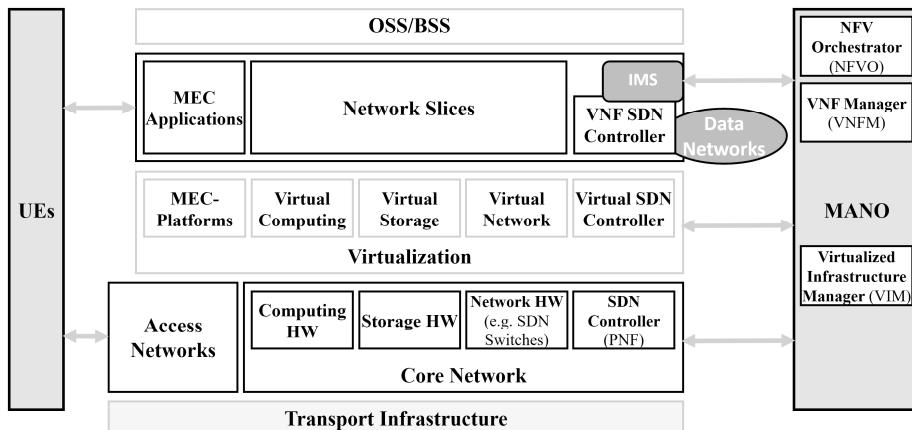
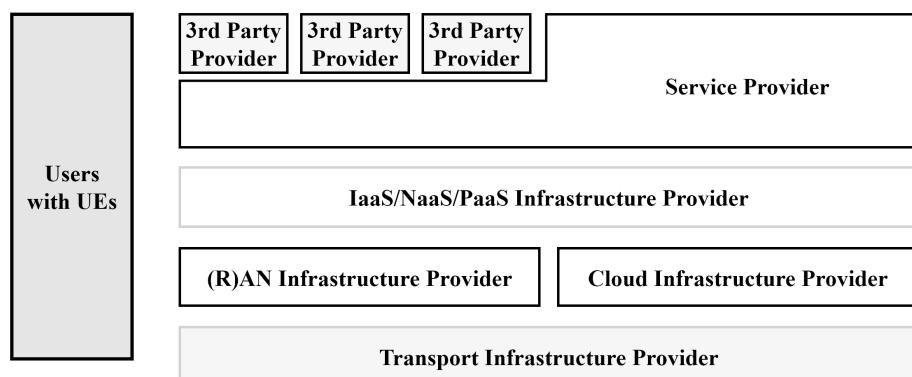


Fig. 10.1: Security-relevant components, subsystems, infrastructures, platforms, and functions of a 5G system

Special security requirements also result from the situation of 5G with several actors (stakeholders) shown in Figure 10.2. Different providers can be responsible for the transport infrastructures, the various access networks, the cloud infrastructures, the IaaS, NaaS, or PaaS platforms provided on them, and the communication services themselves. Besides, 3rd party providers from various industries (verticals) can be integrated as tenants with their own virtual networks, network slices, and/or MEC applications. This includes, for example, a company with its own virtual network for logistics or industry 4.0 applications. A secure operation must be guaranteed for all these providers and tenants [88].



IaaS = Infrastructure as a Service
PaaS = Platform as a Service

NaaS = Network as a Service

Fig. 10.2: 5G system with multiple actors [88]

To make the threat situation for a 5G system even more conscious and illustrative, Figure 10.3 shows possible threats in the different areas of a 5G network. The attacks and threats extend from end devices such as smartphones and IoT devices in, e.g., Industry 4.0 or Smart City environments (Connected World) via the access networks and the 5G core network to the data centers for the 5G applications or the interfaces to the Internet.

Threats on the side of the smartphones are caused by:

- Malicious programs (malware), e.g., for misuse of the operating system
- Spyware, i.e., programs with which information is collected and passed on without the permission and knowledge of the user
- Malicious apps
- Ransomware
- A botnet with malicious programs installed on several or many smartphones.

Specific threats arise in the area of IoT subnets and systems:

- Through botnets of hijacked IoT devices
- Zero-day attacks on previously unknown vulnerabilities, also by combining several attack types
- Attacks with the aim of destroying system-critical infrastructure components, as in the case of the self-replicating Stuxnet worm
- Distributed Denial of Service attacks (DDoS).

The same applies to the threat situation in a smart city, smart buildings and homes, and V2X applications (Connected World).

According to Figure 10.3, a very relevant example of a threat in (R)AN is a possible man-in-the-middle attack.

The 5G core network is particularly vulnerable to

- DDoS attacks and
- Advanced Persistent Threats (APTs). These are complex attack scenarios executed in several stages, using and combining several complex and disguised mechanisms. They are challenging to detect and, in case of failure, are further developed for a new attack. APTs are the next generation of security threats.

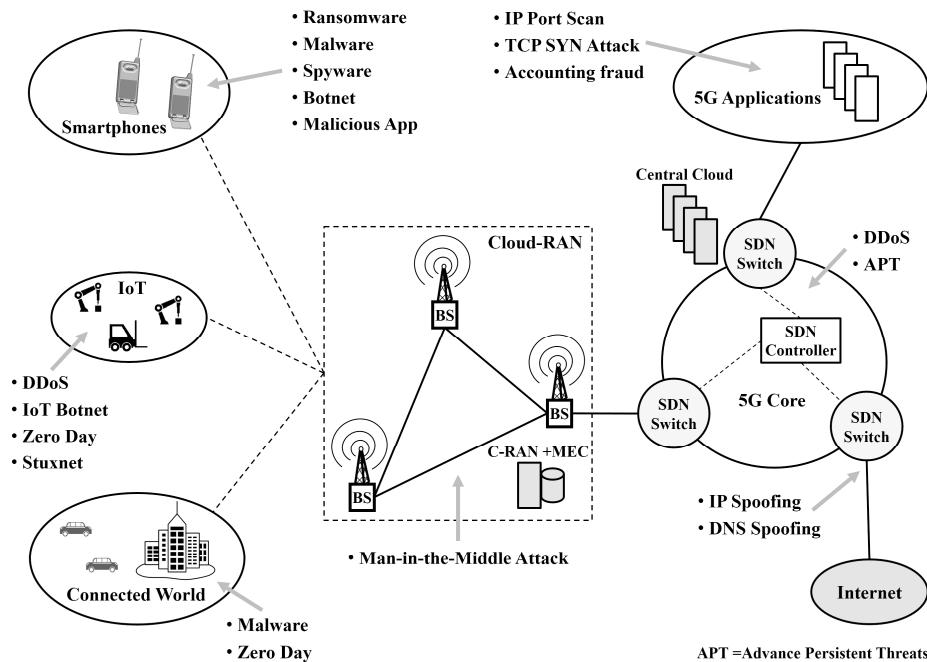


Fig. 10.3: Examples of threats on a 5G network [125]

As far as 5G applications hosted on servers are concerned, threats include

- Port scanning
- TCP SYN flooding attacks
- Accounting fraud attempts.

Attacks at the transition between 5GC and the Internet can be carried out by:

- IP spoofing with fake IP source addresses
- DNS spoofing via forged DNS entries [125; 118].

All in all, the previous considerations and reflections on IT security in a 5G system result in the fact that, on the one hand, we have to provide the usual security functions as in any complex communication network. On the other hand, the security of the central and edge cloud infrastructures used as a result of massive virtualization must be guaranteed. Besides, 3GPP has standardized a 5G mobile network-specific security architecture for 5G as the third pillar with TS 33.501 [45]. These three security pillars shown in Figure 10.4 together form the security framework for a 5G system [159]. We discuss each of these three security areas in the following three sections.

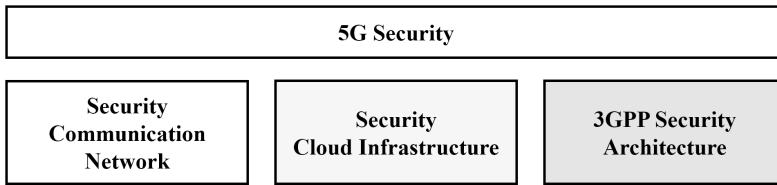


Fig. 10.4: Three pillar model of the 5G security framework [159]

10.1 Security for the Communication Network

According to the first pillar of 5G security, which applies to any telecommunications network, a secure network operation must be ensured. A corresponding catalog of the German Federal Network Agency (Bundesnetzagentur) based on § 109 of the Telecommunications Act (TKG, Telekommunikationsgesetz) [84] describes the requirements. This document, which is generally valid for telecommunications networks, contains the requirements that go beyond the general specifications and are specially formulated for telecommunications providers with IP infrastructure:

- Use of encryption techniques for data and the transport of data
- Measures to protect against DoS and DDoS attacks
- Protection against IP spoofing
- Disabling unused services
- Use of packet filters and adaptive controls (mitigation devices) to limit damage
- Detection of botnets
- Prevention of manipulation of BGP-based inter-domain routes
- Analysis of traffic to detect attacks or errors and to take appropriate protective measures
- Monitoring infrastructure to identify and prevent threats continuously
- Recording of the network management activities
- Logging of target configurations of network components
- Regular target-actual comparison of the configuration data
- Regular testing of network components for compliant behavior
- Also, monitor the network concerning infected systems of customers
- Cooperation in the event of cross-provider disruptions
- Cooperation with anti-malware vendors
- Authentication
- Encrypted transmission of VoIP data
- Correct call number display transmission
- If possible, perform monitoring with SBC for VoIP to detect and prevent TDoS attacks (Telephone Denial of Service) with automated mass calls
- Protection of DNS services.

Also, public telecommunications networks with increased risk potential – and this includes mobile networks – are subject to additional security requirements [84]:

- IT security certification of critical components by a neutral testing agency, in Germany, the BSI (Federal Office for Information Security, Bundesamt für Sicherheit in der Informationstechnik) [227]. To this purpose, a list of critical functions is drawn up and updated, in Germany by the BNetzA and the BSI. Critical components must be identified and notified by the network operator.
- The trustworthiness of manufacturers or suppliers of critical components must be investigated and demonstrated by an appropriate comprehensive and, if necessary, updated source declaration.
- The network operator must verify the integrity of the purchased components throughout their life cycle, from delivery to decommissioning.
- A security monitoring system must be implemented for ongoing operations to identify and prevent threats on an ongoing basis. This includes all critical components as well as the transfer of personal data to contractual partners.
- Security concept for the cryptographic mechanisms and key management
- The professional and personal qualifications of the operating personnel must be ensured.
- Minimization of risks due to technical compromising of critical components, e.g., through redundancies
- For the core network, the transport network, and the (radio) access network, components or systems from at least two different independent vendors must be used. Networks must be designed topologically in such a way that diversity is provided for critical network functions and network elements, i.e., those that require special protection. This could be supported by the application of open standards, such as Open RAN (see Section 7.3).

It should be noted that an analysis of security risks was carried out for Open RAN, considering the protection goals of confidentiality, integrity, accountability, availability, and privacy. The result was that a large number of the interfaces and components specified for O-RAN pose medium to high security risks. This was attributed to the fact that the O-RAN specifications did not follow the security/privacy by design/default paradigm and did not consider the principles of multilateral security (minimum trustworthiness assumptions regarding all stakeholders) [228].

To ensure secure operation according to the above-mentioned requirements, appropriate network protection measures are necessary. In addition to organizational and administrative activities, this includes technical implementation. The separation into security zones must be mentioned here, hence using firewalls and packet filters, application layer gateways, session border controllers, and intrusion detection and intrusion prevention systems. Besides, signaling, user data, and network management must be provided using encryption for secure communication on different layers. The access of end users to the services and operating personnel to

the network components is regulated via authentication and access management based on AAA systems (Authentication, Authorization and Accounting). A redundant system design increases the degree of system availability [93; 118].

The already high complexity of an IP-based telecommunications network becomes even more significant in a 5G system due to the use of NFV with virtual network functions, network slices, and SDN in the transport network. Such a complex and dynamically changing network requires a certain degree of automation for a secure operation to guarantee security. It is not enough to automate the processes. Machine learning and artificial intelligence (AI) methods are needed to be able to adapt to changing network configurations and traffic scenarios in terms of security in order to recognize possible security problems and threats quickly and to react to them as autonomously as possible [159].

10.2 Security in the Cloud Infrastructure

The second pillar of 5G security in Figure 10.4 must provide a secure cloud infrastructure. In this context, we should mention NFV, SDN, the 5G network slices, and the central and edge cloud environments.

An NFV framework (see Section 3.1) for a 5G system, as shown in Figure 10.5, offers a wide range of attack potential and challenges concerning IT security [184]:

- Dependencies on the hypervisor used: There may be security gaps in the software here, affecting the entire NFVI or NFV framework. Therefore a careful patch handling and the application of appropriate encryption is necessary.
- Elastic network boundaries: In an NFV environment, virtual and physical functions are combined, the boundaries between the two worlds are changing. This makes it challenging to provide adequate security functions.
- Dynamic behavior: NFV provides an agile environment in which functions and network topology are constantly changing dynamically. This requires a corresponding dynamic adaptation of the security functions.
- Integration of the security services: In the flexible and dynamically changing NFV environment, the security functions must be continuously and appropriately embedded in the service chains.
- Stateful versus stateless inspection: Until now, IT security has been considered to be better with stateful than stateless security measures. Implementing this at NFV can be difficult due to the dynamic changes and increased complexity.
- Scalability of available resources: The use of deep packet inspection, e.g., by next generation firewalls, is resource-intensive and not well scalable in an NFV environment.

Aggravating factors, according to [160], are:

- Attacks can be carried out via the interfaces between MANO and NFVI or the VNFs.
- Besides, each network function must be protected in its complete lifecycle, as shown in Figure 10.6.
- Furthermore, different tenants' resources and functions must be isolated from each other [160].

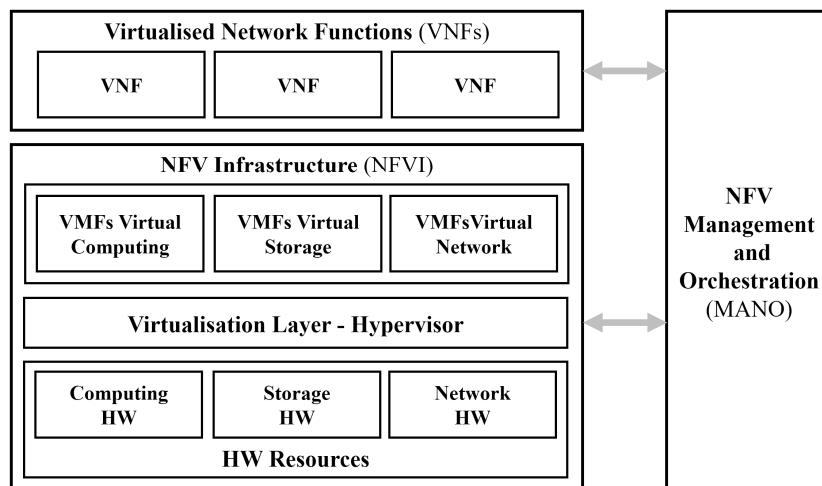


Fig. 10.5: NFV framework

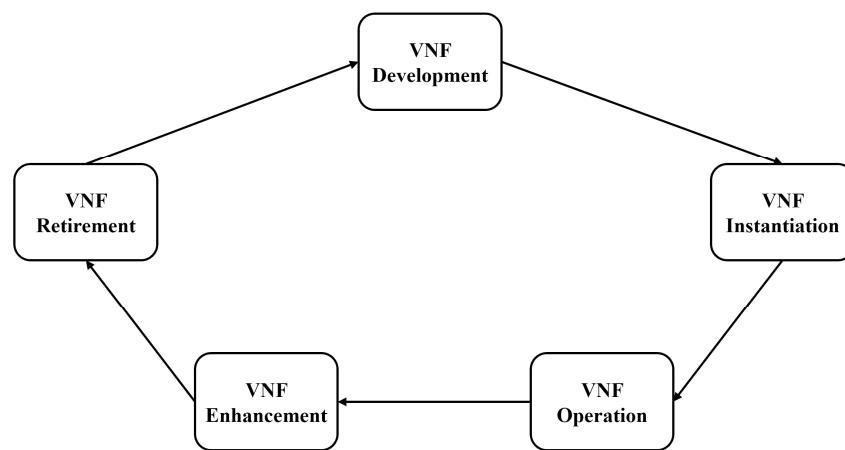


Fig. 10.6: VNF lifecycle [160]

Threats or attack scenarios in the context of NFV [160] include a possible loss of:

- The (5G system) availability through flooding, e.g., as a result of a DDoS attack
- The confidentiality of information through eavesdropping and data leaks
- Data integrity through man-in-the-middle attacks, the takeover of network elements using malware, and unauthorized modification of configuration data.

Table 10.1 summarizes essential NFV security technologies based on the results in [160].

Tab. 10.1: NFV security technologies [160]

Layer	Security measures
Management	Secure APIs
Data	Encryption, metadata security
VNFs	Comprehensive trust measures
Operating system	Secured boot process, hardening, regular patching
Hypervisor	Secured boot process, hardening, regular patching
Computer platform	Hardware supports virtualization, UEFI (Unified Extensible Firmware Interface), TPM (Trusted Platform Module), HSM (Hardware Security Module)

SDN (see Section 3.2) provides the transport infrastructure for the NFV-based network. Figure 10.7 makes it directly apparent that there are threats to IT security at each of the three levels – data plane, control plane, and application plane – and at the interfaces between them – the SBI (South Bound Interface) and the NBI (North-bound Interface) [184; 166]:

- Data Plane: If attackers can manipulate flow tables by accessing one or more SDN switches, they can modify data packets, bypass security network functions (e.g., a firewall), or forward packets to targets defined by them. Also, flooding with packets from previously unknown flows can affect the performance of the SDN network.
- Control Plane: Unauthorized access to the control plane and thus to the controller(s), the heart of the SDN architecture, makes it very easy to manipulate the total data traffic via the possible configuration of the SDN switches and the flow tables. This may affect not only the end users' data but also that for network management and orchestration. Therefore, the controller functionality must be protected by redundancy, extensive access control, software against viruses and worms, firewall, intrusion detection, and intrusion prevention systems.
- Application Plane: Successful attacks at this level lead to modified SDN applications, which are then executed in the controller. These could also be security-related applications with correspondingly far-reaching effects on network oper-

ation and communication services. Among other things, a prescribed authentication of the SDN applications at the controller can help here. Besides, the corresponding software must be developed very carefully from a security point of view.

- SBI: On the one hand, this interface could have a damaging effect on the controller, up to and including takeover, and on the other hand, the SDN switches could be manipulated. The most significant measure against this is the encryption of the messages exchanged via the SBI using TLS.
- NBI: The same threats originate from this interface as from a compromised application plane. Also, here adequate protection is provided by encrypted communication.

Although, as outlined above, SDN provides doorways for security threats. However, SDN also provides a very powerful infrastructure for the network-wide and dynamically adaptable provision of security functions such as firewall, packet filter, application layer gateway, intrusion detection and prevention systems, etc. The controller deploys these by configuring the SDN switches based on the corresponding security applications [166].

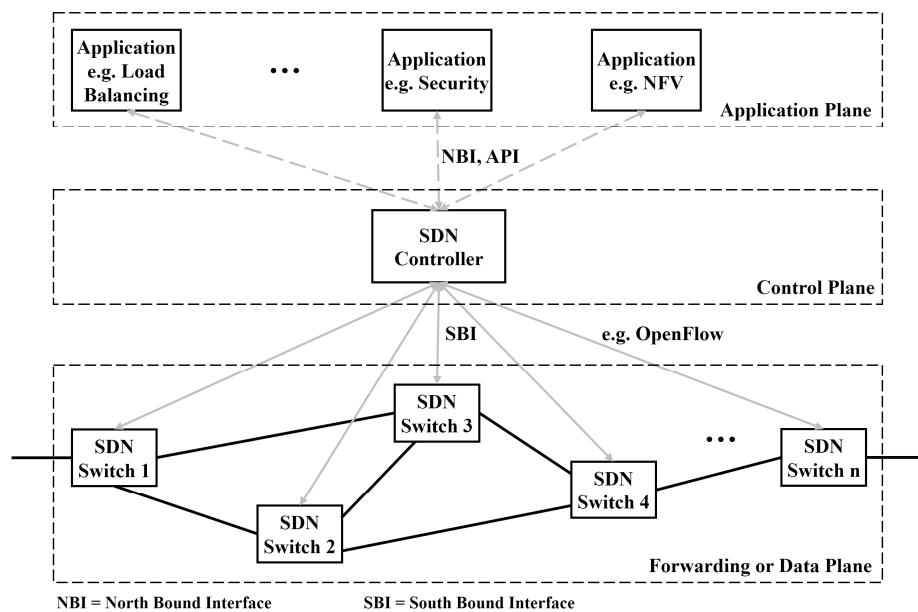


Fig. 10.7: SDN architecture

Network slices (see Section 8.4) based on the Service Based Architecture (see Section 8.3), using NFV and SDN for the very different application scenarios with widely diverging requirements (see Chapter 4), offer the necessary flexibility to meet all of these requirements with a 5G system. Figure 10.8 shows a corresponding example system.

Concerning security, the following aspects must be taken into account [153; 116]:

- Protection of the interfaces and functions of the network slices
- Protection against fraudulent network slice selection
- Prevent unauthorized access to a network slice instance
- Use of different security protocols and policies in different network slices
- Use of different authentication and authorization procedures of tenants of different slices
- Protection against DoS attacks against resources shared by multiple slices
- Prevent attacks from other slices using the same hardware
- The security functions take into account that virtual and physical network functions can be combined within a slice.
- Isolate two network slices, even if the same UE is connected to both at the same time. Under no circumstances may another network slice be reached from an already compromised slice. This also applies to the resources used by both slices.

In order to meet these requirements, the security measures and functions already mentioned in network operation, NFV and SDN are used.

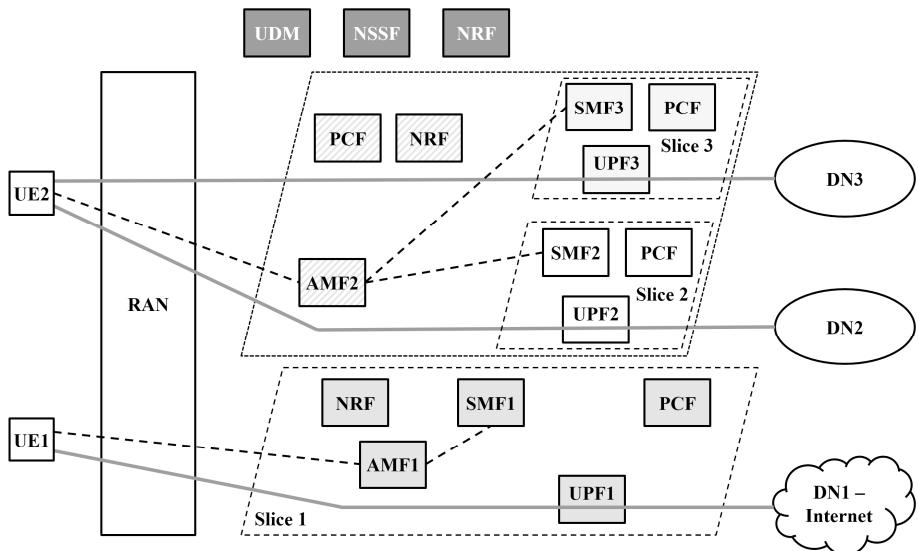


Fig. 10.8: Network slices

Finally, for the area of cloud infrastructure in a 5G system, we will discuss security in central and edge cloud environments (see Section 3.1), illustrated in Figure 10.9. There is also a long list of possible threats [184; 116]:

- Data breach by theft, publication, and/or misuse of protected sensitive or confidential data
- Inadequate identity and access management due to insufficient scalability, weak authentication methods, and poor key and certificate management
- Insecure interfaces and APIs for administration, management, orchestration, and user access
- Weaknesses in system and application software due to bugs
- Taking over the accounts of users by stealing their login information as a result of phishing, fraud, or exploitation of bugs
- Malicious insiders, i.e., current or former employees of the cloud infrastructure operator
- Advanced Persistent Threats (APTs), i.e., professional, targeted attacks over an extended period of time
- Data loss due to unintentional deletion, fire, water damage, or natural disasters
- Misuse of cloud services such as IaaS and PaaS and cloud resources due to inadequate security measures
- DoS attacks on the cloud resources
- Technical problems are caused by sharing the same resources for IaaS and PaaS with multiple tenants.
- Insufficient security of the physical IT infrastructure.

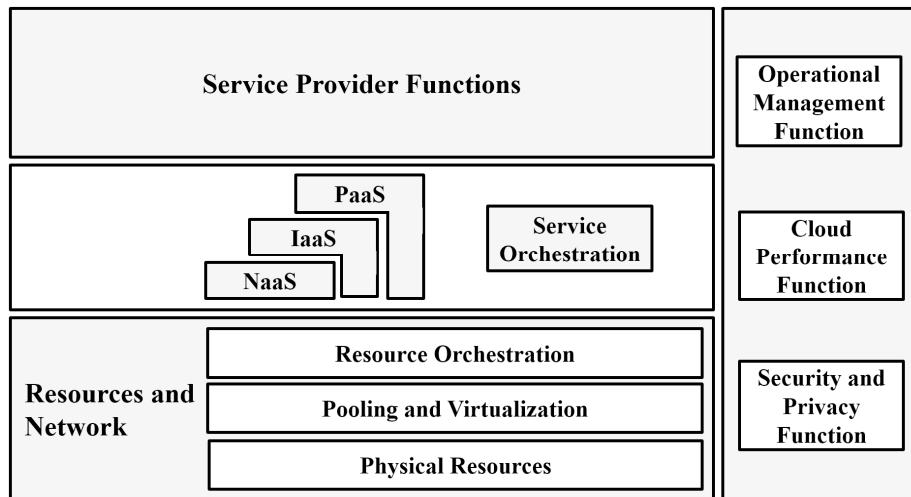


Fig. 10.9: Cloud environment

Adequate security measures for a cloud environment, especially concerning 5G, include [166]:

- Securing the hardware
- Network traffic inspection
- Encrypted communication via the interfaces and APIs
- Strong authentication and access control procedures
- Careful selection of employees as well as transparency and traceability in daily work
- Working according to best practice, monitoring security measures, applying patches and security improvements, conducting vulnerability analyses
- Encryption of data during storage and transport, as well as continuous analysis of data protection, strong mechanisms for key generation
- Two or multi-factor authentication and proactive monitoring of unauthorized activities
- Disclosure of logs and data, infrastructure details such as patch situation and firewalls, as well as information on monitoring and alarms for the tenant, the service provider, by the cloud provider.

For an edge cloud or MEC, there are additional requirements, according to [153]. Since neither signaling nor user data is routed over the core network in MEC use cases, all data required for cost charging must be stored reliably and securely in the edge area. It should be noted that this area of a 5G network is more vulnerable to attacks. Since a 3rd party MEC application may use the same hardware as the network functions of the network operator, it must be ensured that the latter is not affected. Also, the functions of the NFs must not be disturbed by higher bit rate requirements of the MEC application, even as a result of an attack. Furthermore, security data to be stored in the edge area must be specially protected. Besides, extreme latency requirements must not be played off against security.

10.3 3GPP Security Architecture for 5G

The mobile-specific 3GPP security architecture, according to TS 33.501 [45], represents the third pillar of 5G Security in Figure 10.4. The requirements and procedures specified in [45] are based on the security architecture shown in Figure 10.10. It is structured into six security domains, briefly described below:

- Network Access Security (I): This includes the security functions that enable a UE or ME (Mobile Equipment) to authenticate network services and access them securely, both via 3GPP and non-3GPP interfaces in the AN (Access Network). To achieve this, they exchange messages with the Serving Network (SN) via the AN and use the Public Key Infrastructure (PKI) with the keys stored in the USIM

(Universal Subscriber Identity Module) and Home Environment (HE, Home Network).

- Network Domain Security (II): Includes the security functions that enable network nodes to securely exchange signaling messages and user data.
- User Domain Security (III): Protects users' access to mobile devices and services. It also includes hardware-based security mechanisms.
- Application Domain Security (IV): The security functions located here ensure secure communication of applications, both on the user and the provider side.
- SBA Domain Security (V): This ensures security in the SBA with the NFs and their interfaces. Roaming between the home network HE and the visited network SN is also taken into account.
- Visibility and Configurability of Security (VI): These functions, not shown in Figure 10.10, enable users to be informed about the operating status of the security measures and to request additional security functions if necessary.

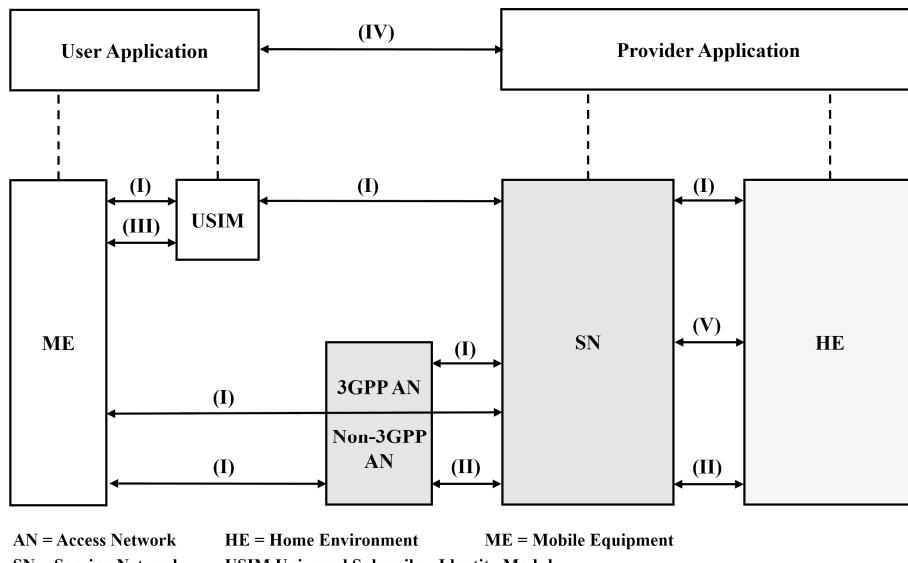


Fig. 10.10: 3GPP security architecture [45]

The security requirements for a 5G system can also be extracted from TS 33.501 [45], summarized in the following.

In general, a UE must be protected against a bidding-down attack (offering a downgrade). In such a case, an attempt is made to deceive the UE so that the UE itself or the network does not support the usual security functions at all and must

therefore communicate with no or reduced security measures. By implementing this requirement, man-in-the-middle attacks can be prevented, for example.

Concerning authentication and authorization, the following requirements are valid:

- The Serving Network (SN) must authenticate the AKA procedure (Authentication and Key Agreement) between UE (ME) and network.
- The UE must authenticate the SN.
- The SN authorizes the UE based on the user profile received from the Home Network, HE.
- The HE authorizes the SN; this is communicated to the UE connected to the SN.
- The SN authorizes the desired AN. As a result, the UE receives the assurance that the AN used by it provides the chosen services with the required security.
- Emergency calls are possible without authentication [45; 153].

Table 10.2 summarizes security requirements for 5G network elements and 5G network functions (see Chapter 8) [45].

Tab. 10.2: Security requirements for 5G network elements and 5G network functions [45]

5G network elements or 5G network functions	Security requirements
UE	<ul style="list-style-type: none"> – Encryption of signaling and user data between UE and gNB for reasons of confidentiality – Ensuring the data integrity for signaling and user data between UE and gNB – 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 SUCI (Subscription Concealed Identifier)
gNB	<ul style="list-style-type: none"> – Encryption of signaling and user data between UE and gNB for reasons of confidentiality – Ensuring data integrity for signaling and user data between UE and gNB – Authenticating and authorizing a gNB during setup and configuration – Protection of the gNB software – Protection of the keys used and stored in the gNB – Secure processing and storage of user and signaling data – Providing a secure environment for all sensitive data – Secured transmission on the F1 interface when splitting a gNB into CU and DU – Secure transmission on the E1 interface when dividing a CU into CU-CP and CU-UP

5G network elements or 5G network functions	Security requirements
AMF (Access and Mobility Management Function)	<ul style="list-style-type: none"> – Because of the confidentiality encryption of NAS signaling – Ensuring data integrity for NAS signaling – Triggers the primary authentication of the UE via SUCI (Subscription Concealed Identifier)
SEAF (Security Anchor Function)	<ul style="list-style-type: none"> – Provides the authentication function via the AMF in the Serving Network – Supports the primary authentication of the UE
UDM (Unified Data Management)	<ul style="list-style-type: none"> – Long-term keys for authentication and security association must be protected and must not leave the UDM/ARPF environment (Authentication credential Repository and Processing Function). – Provides SIDF service
SIDF (Subscription Identifier Deconcealing Function)	<ul style="list-style-type: none"> – Responsible for resolving the SUPI (Subscription Permanent Identifier, the unique ID on SIM card) from the SUCI
AUSF (Authentication Server Function)	<ul style="list-style-type: none"> – Processes authentication requests for 3GPP and non-3GPP access – Informs about it UDM – Transfers the SUPI to the VPLMN (Visited Public Land Mobile Network) after successful authentication
Core network in general	<ul style="list-style-type: none"> – Creation of trust zones, in any case between different 5G network providers – Secure discovery and registration of NFs in the SBA – Authentication between NF producer and NF consumer – Validation of each received message by NFs – Secure end-to-end connections for the application layer between 5G core networks
NRF (Network Repository Function)	<ul style="list-style-type: none"> – The NRF and service requesting NFs must authenticate each other. – The NRF provides authentication and authorization to NFs for secure communication between themselves.
NEF (Network Exposure Function)	<ul style="list-style-type: none"> – Ensures confidentiality and data integrity between NEF and AF (Application Function) – Mutual authentication – Does not communicate information about network slice or SUPI to the exterior
SCP (Service Communication Proxy)	<ul style="list-style-type: none"> – Provides routing and message forwarding to destination NFs – Authentication between SCP and the NFs and between SCPs – All communication between SCP and NFs and between SCPs must be confidential, integer, and protected against replay attacks (using data previously recorded by the attacker)
SEPP (Security Edge Protection Proxy)	<ul style="list-style-type: none"> – Protects communication between NFs in different PLMNs (Public Land Mobile Network) – Mutual authentication with corresponding SEPP – Hiding of the own SBA (topology hiding) – Application Layer Gateway functionality

5G network elements or 5G network functions	Security requirements
NSSAAF (Network Slice Specific Authentication and Authorization Function)	<ul style="list-style-type: none"> – Processes slice-specific authentication and authorization requests from the AMF – Communicates with the corresponding AAA server and provides protocol conversion

Based on the requirements mentioned above, we can now provide an overview of the 3GPP security architecture implementation. Figure 8.21, from the description of the SBA approach to roaming in Section 8.3, is used here. If we take all network functions with security tasks from Table 10.2 and show them in grey in Figure 8.21, we will get Figure 10.11, which we will use for an introduction to the security procedures in a 5G system [159].

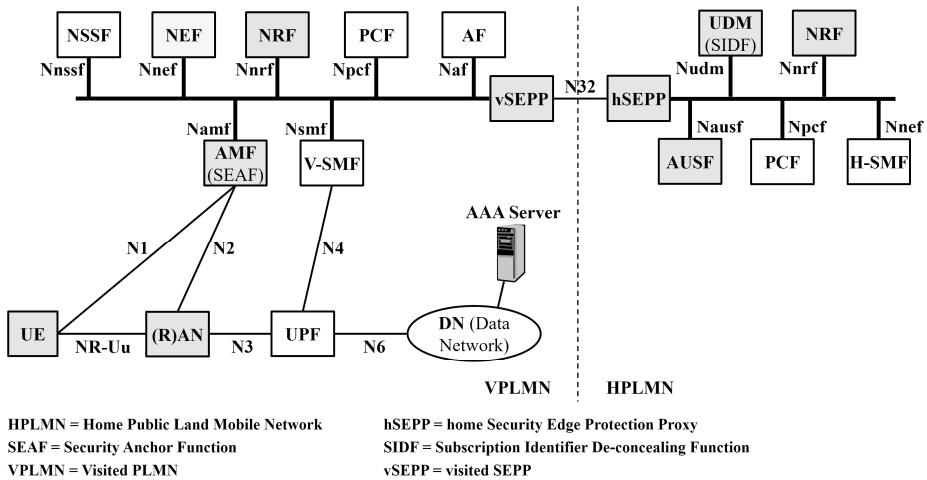


Fig. 10.11: 5G systems with SBA and security features for the case of roaming [37]

The UDM (Unified Data Management) in the home network (HPLMN) enables access to the subscriber profile and thus a cryptographic key per user. This shared key is also stored in the mobile device, the UE, on the SIM card. If there is an authentication between UE and network, the AUSF (Authentication Server Function) retrieves the necessary information in the UDM and derives a key valid only for this one session to secure the signaling between UE and the HPLMN, without access to the visited network, the VPLMN. A second key is generated for use by the VPLMN. These

temporary keys are generated based on the user-specific shared key mentioned above not only in the network but also in the UE. The only information to be exchanged via the radio interface is a random value generated by the respective network, valid only for this session, which is used for generating the session key [159]. According to [45], there is a hierarchy of keys derived from each other.

The AMF (Access and Mobility Management Function) handles the part of the authentication procedure for the VPLMN. Based on the session key received from the HPLMN, a secure signaling connection is established between UE and the AMF [159].

As shown in Figure 10.11, signaling traffic between interconnected 5G core networks is always secured via SEPP network elements (Security Edge Protection Proxy) at the border between the two 5G networks, in the case of roaming one in the visited (vSEPP) and one in the home network (hSepp). Therefore, the entire authentication process occurs via UE, AMF, and vSEPP in VPLMN, from there in HPLMN via hSEPP with AUSF. The SEPPs use security mechanisms that allow only part of the signaling information to be encrypted. In contrast, other parts are transmitted unencrypted, e.g., integrating a connection network operator or a broker for roaming. However, it is ensured that only network nodes authorized for this purpose are involved [159].

Another network function that is particularly important for security in Service Based Architecture (SBA) is the NRF (Network Repository Function). It not only offers the other NFs the possibility of registering with its functionalities and requesting information on other NFs that may be required, but it also operates as an authorization server. It decides which calls are allowed between which NF instances. For this purpose, an NF consumer must be authorized by the NRF to use the service of a producer NF using the OAuth procedure following RFC 6749 [13]. For this purpose, the NRF checks whether the access is allowed by the configured rules. If yes, the consumer NF receives an authorization token, which is transferred to the producer NF during the also secured HTTP/2 over TLS-based access and checked by the producer NF. Only if the token is valid, the service call is answered positively. Where appropriate, this mechanism may be limited to calls from other networks. In summary, the NRF is a central network element in the security architecture of a 5G system [159].

A 5G base station, the gNB in the (R)AN in Figure 10.11, provides secure communication on the radio interface to the UEs for both signaling and user data. It is based on a key that is generated per session and gNB in the network and the UE and which is also changed during handover to another gNB. It means that there are always two secure connections in terms of signaling, one between UE and gNB, and one between UE and AMF. Besides, all signaling and user data traffic to and from the UPF is transmitted between the gNB and the 5G core network in an IPsec tunnel and is therefore encrypted. IPsec is also frequently used for the secure transmission

of user data between 5G PLMNs. For secure end-to-end transport, the TLS protocol is suitable for many applications [159].

In Table 10.2 and Figure 10.11, the NEF (Network Exposure Function) is also highlighted as security-relevant. It provides mutual authentication with an AF (Application Function) of a 3rd party provider and also authorizes AF accesses to NFs in the SBA using the above-mentioned OAuth procedure. The message exchange between NEF and AF is TLS-secured [45].

For access to the DN (Data Network) in Figure 10.11, after the first authentication between the UE and the 5G network outlined above, a second authentication (Secondary Authentication) can optionally be performed against the included DN to increase security. The EAP procedure (Extensible Authentication Protocol) according to RFC 3748 [8] is used. The UE acts as an EAP client, the AAA server as an EAP server. The authentication method used on top of EAP can be selected according to the circumstances and is determined here by the H-SMF (Home-Session Management Function) as the intermediate EAP authenticator [45].

Furthermore, additional security functions were specified in TS 33.501 [45] for the Cellular IoT application area (see Section 9.4). This includes transmitting user data of short length or SMS as NAS messages (see Section 8.1). UE and AMF ensure integrity protection and encryption using the NAS security context. In addition, an RRC connection between UE and ng-eNB may be temporarily suspended and re-established by the UE with the same or another ng-eNB using the same security context. In addition, the secure transmission of unstructured data (non-IP) is also supported.

11 5G and Environment

In this chapter, we look at the environmental impact of the 5G rollout. This includes the question of the influence of non-ionizing radiation due to radio transmission in Sections 11.1 to 11.4, and in Section 11.5, energy consumption, the need for raw materials, and generally the contribution to sustainability.

11.1 New Issues through 5G Technology

If we deal with the topic “5G and environment” and thus, of course, with the question of the resulting electromagnetic radiation due to the introduction of 5G systems, it is first of all interesting to see what is technically different from 4G.

As mentioned in Section 5.1, the new frequency ranges to be applied should be considered. In the first step, this is in most countries only the range between 3.4 and 3.8 GHz (see Section 5.3). Besides, 5G also uses low-frequency spectral bands that are shared with 4G or become available. The mentioned high-frequency ranges are necessary, since only here the channel bandwidths of up to 100 MHz, which are essential for the required high bit rates, are available. One disadvantage, however, is that these higher-frequency radio signals are more strongly attenuated by obstacles such as walls, windows, trees, etc., and even by the air in the open space, compared to 4G. The use of adaptive antennas can partially compensate for these unfavorable transmission characteristics. Such antennas consist of several individual components (massive MIMO), which are controlled in the case of a pending transmission so that the overall characteristic of the antenna points precisely in the direction of the mobile terminal concerned during this time. The transmission power is briefly concentrated on this particular transmission path by this so-called beam-forming, the higher attenuation is compensated, and interference in the radio cell and from neighboring cells is reduced (see Section 7.1). As far as electromagnetic radiation is concerned, the average exposure in the radio cell is lower but higher for persons in the vicinity of the antenna due to the beam.

In a second step, millimeter waves from 24.25 GHz onward are also used (see Section 5.1). This is necessary for the very high bit rates of up to 20 Gbit/s (see Section 4.3) since channel bandwidths of up to 400 MHz are required, and these are not available in lower frequency ranges. At these frequencies, one has to deal with even worse propagation characteristics. In most cases, line-of-sight will be necessary. The reasonable coverage distance will often be only a few meters. Consequently, the radio cells have only a small size when millimeter waves are used. Therefore, the base station is located relatively close to the user, but the transmission power is also significantly lower than in a macrocell. Besides, when using millimeter waves, even

holding the terminal device by hand can lead to strong shadows and an upregulation of the transmission power.

Cellular mobile radio networks consist of many adjoining and partly overlapping radio cells. Macrocells form the backbone of such an access network with powerful antennas, usually installed on free-standing masts or house roofs. They provide comprehensive coverage of an area and typically have a coverage radius of 200 meters to typically 2, often no more than 5 kilometers. The transmission power is dimensioned in such a way that the radio signals emitted still reach the terminals in buildings, in vehicles, and also at the edge of the cell, but without disturbing the signals in neighboring cells. In areas with very high data traffic, macrocells are supplemented with microcells to increase capacity. Microcells with lower transmission power typically provide coverage of 50 to a maximum of 200 meters outdoors. In addition, so-called picocells with a coverage range of less than 100 meters are used in special situations, e.g., in hospitals or shopping malls inside buildings, and outdoors, e.g., at bus stops. A radio cell that only supplies an office, for example, is called a femtocell. Because of these different types of radio cells, one speaks of a hybrid network. Such an access network architecture is quite normal for mobile networks, even with 4G and 3G. However, the number of micro and even smaller cells increases the higher the frequencies used. This means there is a clear trend towards micro and picocells for 5G, especially in cities, to meet data transmission requirements. In these cases, the base station antennas are correspondingly closer to the users.

Finally, when comparing 5G to 4G, it should be mentioned that the radio transmission technology is basically the same. Both systems use the same modulation method. For this chapter, interesting differences exist in the signaling and control data volume. Here, compared to 4G, with 5G, there are only 20% of corresponding signals. This means that exposure to electromagnetic radiation is significantly reduced at 5G during periods of low traffic [92].

11.2 Electromagnetic Radiation and Health

For radio waves, non-ionizing radiation (NIR), the International Commission on Non-Ionizing Radiation Protection (ICNIRP) [110] has set immission limits. ICNIRP is a non-profit scientific institution that publishes recommendations for NIR limits based on the evaluation of numerous scientific studies, taking into account safety factors. Legislators and authorities often refer to these in turn. For a long time, the ICNIRP Guidelines of 1998 [111] formed the basis, also for NIR at 5G. New guidelines [112] have been available since March 2020, but they essentially confirm the previous recommendations. Compared to the earlier editions, they are more precise for short exposure times of less than 6 minutes and exposure of small body areas of a

few square centimeters. In particular, more attention was paid to frequencies above 6 GHz (see Section 5.1), which will be relevant for 5G in the future [113].

Especially for the USA, in addition to ICNIRP, the Institute of Electrical and Electronics Engineers (IEEE) should be mentioned, which has also developed exposure guidelines and defined exposure limits for frequencies up to 300 GHz [23].

Compliance with the recommended guideline values is intended to protect people from thermal effects, in particular, i.e., the warming of human body tissue by absorption of radiation should be strongly limited. However, the immission limits do not include biological, so-called non-thermal effects in the low-dose range and scientifically unproven long-term effects.

A current and comprehensive summary of this topic can be found in [92; 94]. A working group compiled this report in Switzerland on behalf of the Federal Department of the Environment, Transport, Energy and Communications (Eidgenössisches Departement für Umwelt, Verkehr, Energie und Kommunikation, UVEK). Based on available studies, the association between mobile radio radiation and cancer risk and other health effects was evaluated according to a common scheme. The evidence, i.e., the result, is thereby classified as sufficient, limited, insufficient, or absent.

“In relation to possible health effects of 5G radio technology, there are as yet few studies on cells and animals relating to acute effects. The working group's risk assessment therefore relied on studies conducted in the past on 2G, 3G and 4G technology and which work with frequencies which lie in the same range as those frequencies currently being used for 5G.”

“The working group determined that to date, for the mobile radio frequencies currently in use, no health effects below the guideline values of the international radiation protection commission ICNIRP, on which the immission limit values of the ONIR (Ordinance on Protection from Non-ionising Radiation [142]) are based, have been consistently scientifically proven.”

“However, the question for the working group is whether, with reference to the precautionary principle, there are indications or proven findings for effects below the ICNIRP limit values (and the ONIR immission limit values respectively). The working group assesses the evidence of effects as follows:

- In 2011 the International Agency for Research on Cancer (IARC) classified high-frequency radiation as possibly carcinogenic for humans, on the basis of the results of studies on mobile telephone use, with indications of increased risks for gliomas and tumours of the auditory nerve. Since 2014 two important large animal studies have appeared which give indications of carcinogenic effect for mobile radio radiation. The results of new population-based studies on the connection between mobile telephone use and tumour development have so far been inconsistent. Most investigations carried out to date in several cancer registers indicate no increases in disease rates. Overall, the evidence for a carcinogenic effect is assessed as limited, as it was in 2014.

- On the question of tumours in connection with mobile communication base stations, television and radio transmitters there are still very few studies. A study published in 2014 found no connection between TV and radio transmitter exposure for all cases of child cancer diagnosed in Switzerland between 1985 and 2008. In the case of lower exposure due to transmitter installations, the evidence is judged to be insufficient, as in 2014.
- A study on mice published in 2015 was able to confirm earlier results according to which simultaneous exposure to high-frequency NIR and exposure to a proven carcinogenic substance causes faster tumour growth than with the carcinogenic substance alone. Replication of this tumour promotion could be used as an argument for upgrading the evidence from limited to sufficient. However, the absence of an exposure-response relationship and methodical limitations in the study, as well as the absence of confirmation for a tumour-promoting effect from an epidemiological study, are arguments against upgrading the evidence of co-carcinogenesis. Overall, therefore, the evidence for co-carcinogenesis continues to be assessed as limited.
- There is sufficient evidence of physiological effects on humans in the event of exposure of the brain to radiation intensities within the range of the ICNIRP guidelines for local exposure. Thus a series of experimental studies with test subjects came to the conclusion that exposure within the intensity range of the ICNIRP guideline value for exposure due to a mobile telephone against the head affects brain waves in the at-rest waking state as well as during sleep. However, since the quality of sleep was not impaired as a result, the significance of this effect for health is unclear. Some of these experimental studies also found different effects as a function of modulation, which indicates that in addition to signal strength the signal form of the exposure could play a part. The extent to which the signal characteristic (e.g. modulation) plays a part has still not been adequately systematically evaluated.
- There are hardly any studies on humans in which the entire body is exposed within the range of the ICNIRP whole-body guideline value, corresponding to the immission limit value for mobile radio base stations. In everyday life such exposures practically do not occur, although they are permissible in principle up to the limit value, making observational studies difficult. In epidemiological studies the persons most exposed are exposed to levels significantly lower (approx. 0.2-1 V/m) than the whole-body limit value. A series of new studies from Holland and Switzerland found no link between the occurrence of symptoms and NIR exposure at the place of residence. This indicates that there is no such link (evidence for absence). In these studies (as also in reality) the proportion of persons who are subject to higher exposure compared to the average is very small. The studies are therefore not sufficiently conclusive to assess effects of exposures in the range of the installation limit value and above (evidence insufficient).

- In medical practice there are cases in which patients plausibly attribute their complaints to high NIR exposures in their everyday life. However, proof of such an effect cannot be provided in individual cases. In double-blind, randomised studies no proof of such electromagnetic hypersensitivity could be provided, though predominantly the perception of short-term exposure was investigated. It cannot, however, be excluded that the effects manifest themselves only under certain conditions or exposure situations which are not yet understood. Owing to methodical difficulties with investigation of electromagnetic hypersensitivity, additional research activities are therefore urgently required.
- Many cell studies and animal studies have been carried out. These frequently find biological effects, but the results are not uniform. Thus, for example, there is no consistent pattern with regard to exposure/effect relations or to the question of which cells are particularly sensitive. Since these studies include a multitude of biological systems and the corresponding expertise was not represented in the working group, they were not assessed in depth. Accordingly, there is also no evaluation of evidence.
- There are already a few cell studies and animal studies on exposures within the 30 to 65 GHz range (millimetre waves). However, the results are not sufficiently robust for an evaluation of the evidence.” [92; 94]

When building 5G, frequencies between 3.4 and 3.8 GHz are increasingly used, which have a similar absorption behavior in the human body as the frequency bands already used in mobile communications. However, compared to lower frequencies, the energy is absorbed less in the internal organs of the body. About 95 percent of the energy is absorbed in the skin and up to 2 cm below. With millimeter waves above 24 GHz, the waves penetrate the tissue even less. The skin and eyes are mainly affected [92; 79]. In this regard, various metastudies for these high frequency ranges have so far judged the results to be of little significance concerning possible health effects, and further research is seen to be needed [164; 236; 92].

“Health effects can never be scientifically excluded with absolute certainty. The working group therefore also described the potential effects for which further research is indicated.” [94]

According to [92], these are, among others:

- “There are already many studies on the biological effects of HF-NIR (High Frequency-Non-Ionizing Radiation) below 6 GHz, but even less on millimeter waves. Studies should clarify whether these frequencies have other biological effects.
- It has not yet been fully clarified how relevant the signal characteristics (e.g., modulation) are in all frequency ranges used by mobile radio.
- With the higher frequency and thus decreasing wavelength, new demands are placed on dosimetry. An image of the skin with its layers as close to reality as possible is indispensable.

- National health study
- In addition to population studies, an in-depth investigation of persons who attribute health problems to NIR is also conceivable.
- Since practically the entire population uses a mobile phone, we can expect that any tumor risk would have to be reflected in an increase with a certain latency period. It is therefore proposed to establish monitoring of brain tumors.”

In [79], the Federal Office for Radiation Protection (BfS, Bundesamt für Strahlenschutz) in Germany [78] evaluates mobile communications, including 5G, from the radiation protection point of view. In the following, excerpts are quoted in addition to [92] about possible adverse health effects.

„Biological effects of HF fields on cell functions, genetic material, gene expression, and the neurophysiology of the visual and auditory system have, to some extent, been investigated in replication studies. A review of the „melatonin hypothesis“ and the hypothesis regarding a demodulation of HF fields failed to confirm any of the postulated effects.“

“Several volunteer studies have investigated the possible influence of HF fields from mobile phones on brain performance and sleep. Minimal EEG changes have been observed but they are not considered to be relevant to health. Impairments of well-being (headaches, dizziness, etc.) that some people attribute to HF fields can lead to a significant reduction in the quality of life of affected individuals. Provocation studies under laboratory conditions and studies in real-life conditions have uniformly failed to support causal relationship between HF fields from mobile radio and impairments of well-being.”

The question of carcinogenic effects of mobile radio fields „has been intensively explored in epidemiological studies and animal experiments. With regard to mobile phone usage, the majority of epidemiological studies show no increase in the risk of tumours either in general or in the head, throat and neck region. However, the epidemiological studies published so far do not allow a conclusive statement regarding the risk of brain tumours in „heavy users“ of mobile communications technology or – given the comparatively short period of use so far – the risk of cancer in the event of use over a period of more than 15 years. In animal experiments involving lifelong exposure over several generations, no reliable relationship has been identified between cases of cancer and exposure to different mobile radio standards at levels below the limit values“.

„Animal studies have investigated whether mobile radio fields might potentially promote the growth of existing tumours. Carcinogenic substances were used to produce tumours in mice or rats, and the animals were then exposed to various mobile radio fields. Although no such effects were identified in a number of animal models, exposure to mobile radio (3G/UMTS) did promote the growth of tumours in a specific strain of mice – primarily in the lungs and liver. As the exposure of people (to UMTS signals) results in completely different field distributions inside the body (the

fields do not reach the lungs or liver, for example), this result is not directly applicable to humans. Considering these study results as a whole, the BfS therefore assumes that the growth of tumours is not promoted in humans.“

„Other theoretically possible consequences of longterm exposure have been examined, primarily in animal studies. These studies have not provided reliable evidence that mobile radio fields of different standards/generations affect the permeability of the blood-brain barrier, the occurrence of tinnitus, male fertility, the immun system, nerve cells or the stress response. Likewise, multigenerational studies have not identified an influence on learning or brain performance in offspring.“

„Higher frequency bands (> 20 GHz) have a shorter range, and the corresponding HF fields are absorbed very close to the surface of the body (the penetration depth of the fields is frequencydependent and is 1 mm or less from 20 GHz upwards). Direct effects on internal organs are therefore not to be expected. If these millimetre waves had any effect at all, they would primarily affect the skin and eyes. As this topic is still a subject of relatively little research, the BfS is commissioning research projects in this frequency range.“

In summary, the BfS's assessment is: „Adherence with the limit values provides protection against all scientifically proven health effects due to HF fields, but there are still uncertainties when it comes to the risk assessment of intensive and long-term mobile phone use, as well as regarding the use of millimetre waves.“ [79]

As mentioned above, the IARC of the WHO (World Health Organization) classified radiofrequency radiation as possibly carcinogenic for humans. This is also followed by the meta-study [236] from the research service of the European Parliament, whereby in [236], the effects on human reproduction/development are also seen more critically. A meta-study is currently being conducted on behalf of the WHO, focusing on the effects of exposure to radiofrequency electromagnetic fields (RF-EMF) on male fertility and pregnancy and birth outcomes [235]. This study is intended to meet the highest scientific standards. Results were not yet available in late summer 2023.

11.3 Exposure and Limit Values

To characterize the exposure of the population to non-ionizing radiation (NIR), different factors are relevant [92]:

- The emissions denote the transmission power of a source in W (Watt). The Effective Radiated Power (ERP) of an antenna is also used frequently, i.e., the power fed into the antenna multiplied by the antenna gain.
- The distribution of NIR in the environment is represented as immissions, as electric field strength in V/m (volts per meter), or as power flux density in W/m².
- Exposure refers to NIR at the location where a person is present, quantified in V/m, or W/m².

- The dose refers to the NIR absorbed by the body, the Specific Absorption Rate (SAR) in W/kg. If it is absorbed over a certain time, it is called a cumulative dose. The cumulative dose is obtained by multiplying the SAR value by the time duration. It is quantified in J (Joule) per kg body weight per day.

Emissions can be from distant sources such as base stations or other users' mobile devices and/or near-body sources such as smartphones. Table 11.1 shows the basic differences regarding exposures. The immissions differ in frequency (e.g., 2,6 GHz (LTE) or 3,5 GHz (NR)), intensity (e.g., base station or terminal), temporal pattern (e.g., strongly changing signal of a base station or a rather continuous signal of a broadcasting station) and signal shape (e.g., OFDM for LTE or NR or relatively sinusoidal signal for broadcasting). The immissions depend strongly on the distance to the emitter. An active smartphone at the ear leads to significantly higher immissions for the user than the base station 2 km away, even though it transmits much higher power.

Tab. 11.1: Exposure from a base station and terminal device in comparison [92]

Base station of a macrocell	Mobile end device
Relatively strong transmitter	Weak transmitter
Larger distance to persons	Very small to a small distance to the body
Low absorbed power	Very high locally absorbed power
Exposure permanent, but fluctuating throughout the day	Exposure only during a phone call or data transmission
Large area radiation with exposure of all persons in the vicinity	Exposure mainly to the user and persons in the immediate vicinity

Which exposure level is biologically particularly relevant, the average immission, or maximum value, or a level exceeded in a certain time, or a special signal form, etc., is not yet known. The maximum load and average immissions can be influenced by network architecture, as explained in Section 11.4.

The use of a smartphone leads to emissions close to the body, resulting in exposure of the head or hand of a person. With increasing distance, the exposure decreases rapidly, e.g., by using a headset. The higher the transmitting power, the higher the exposure. The exposure depends on the terminal device, more precisely on its preferably low SAR value, on the mobile radio technology (NR and LTE better than GSM), and above all, on the connection quality between the terminal device and base station. Due to the power control, short distances and lack of obstacles lead to the lowest exposures.

Mobile phones and other terminal devices have, in most cases, a much lower transmission power than a base station. However, the exposure of humans to the terminal during a telephone call or data transmission is usually much higher than that of the most powerful base station. That's because the terminal device is often only a few millimeters from the head or a few centimeters from the body, while the antenna of a base station is rarely positioned closer than a few meters. Due to the larger distance to the base station, the whole body is uniformly exposed to its radiation. In contrast, the terminal device mainly irradiates the head or hand [92].

Compliance with the legally defined immission limits is intended to protect against scientifically proven health effects, i.e., among other things, heating of body tissue by more than 1 °C within 30 minutes. These limits must be observed wherever people can be present, even for short periods. In Germany, they are between 39 V/m (around 800 MHz) and 61 V/m (around 2.6 GHz (LTE) and 3.6 GHz (NR)) [80; 53], in Switzerland between 36 and 61 V/m [142; 92].

In Switzerland, the stricter plant limit values also apply. They are below the immission limit values and were introduced based on the so-called precautionary principle to consider a precautionary measure any as yet unknown effects that could be harmful to humans. A mobile radio installation (possibly with several, even adaptive antennas of different operators at the same site) may, at maximum transmission power, expose places where people regularly spend long periods of time (e.g., schools, children's playgrounds, hospitals, apartments) to a maximum of about 1/10 of the immission limit value (i.e., between 4 V/m (up to 900 MHz) and 6 V/m (from 1.8 GHz)). However, this only applies to macrocells, not transmitters with 6 W or less transmitting power [142; 92].

11.4 Influences of the Network Architecture

The immissions caused by a base station are influenced by its transmission power, the transmission direction or the antenna pattern (also considering adaptive antennas), the distance to the antenna, the attenuation by free space, and especially by obstacles as well as the amount of transmitted data. These immission values can be minimized by:

- The cell size is small.
- The data rate and thus, the required bandwidth is as low as possible.
- The base station is close to the users.
- There are as few obstacles as possible between the base station and the user.
- As few base stations as possible per site
- Using beamforming.

The latter is because the exposure averaged over the area is expected to be lower than with conventional antennas due to directed transmission to the requesting terminal.

Otherwise, the optimization mentioned above ensures that the transmission power of the base stations often added over the systems of several operators and mobile network generations (5G, 4G, 3G, and 2G), can be kept relatively low. Besides, the end devices then have better connection qualities and, therefore, also require less transmission power, reducing exposure. All in all, these are arguments against macro and for microcells. For mobile communication in buildings, pico or femtocells would be preferable to micro or macrocells operated outside. If the latter is used, the transmission power of the base station and terminal device must be correspondingly higher due to the necessary wall penetration. Consequently, emissions are also higher [92]. It should also be noted that buildings that meet high or the highest energy standards strongly attenuate mobile radio radiation, e.g., due to triple-glazed windows, to the extreme that no indoor coverage is possible in the building without additional antennas.

Figure 11.1 shows network architectures with different cell types.

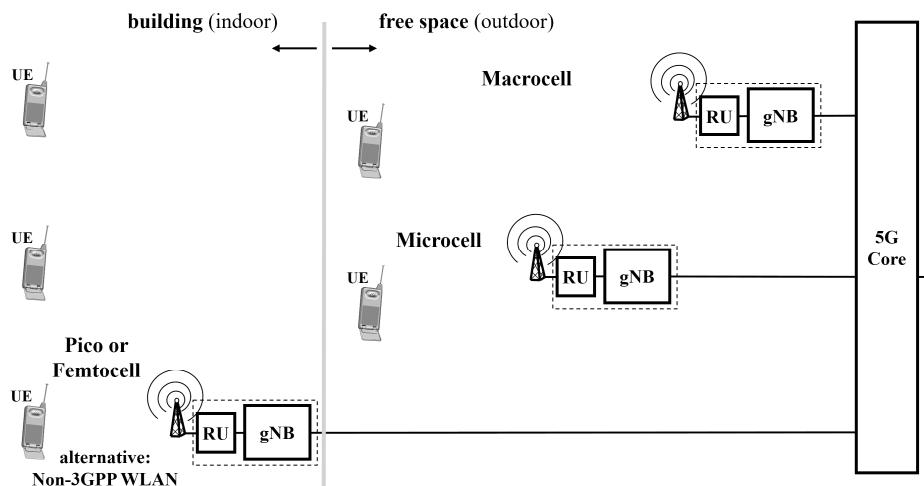


Fig. 11.1: Influence of the network architecture with different types of radio cells on NIR immissions

In the case of the comparatively large macrocell, the UEs can be relatively far away from the base station, and there can be massive obstacles in the transmission path, e.g., concrete walls of buildings or the building insulation mentioned above. In these situations, the base station and a mobile terminal must increase their transmission power, and the exposures are accordingly high. In the case of a microcell, which has a significantly lower maximum transmission power anyway, this applies

only to a limited extent. However, even here, an indoor UE must work with highly regulated transmission power and generates correspondingly high immissions. These disadvantages can largely be avoided by cleverly placing base stations for pico- or femtocells in buildings. Exposure can be minimized by obstacle-free communication over only short distances. As an alternative to a 3GPP pico or femtocell, untrusted or trusted non-3GPP access via WLAN with VoWifi (see Section 2.6) could also be used indoors. In these network scenarios, it would also be sufficient to install not the entire base station in the building but only the RU, i.e., the radio transceiver with an amplifier, and operate the actual gNB remotely. In addition, for the reasons mentioned above, it is already common practice today in larger buildings to supply the rooms via an in-house antenna infrastructure and to house the radio transmission technology or the gNB in the basement, for example.

In summary, NIR immissions and exposures could be minimized by appropriate intelligent network architecture, as shown in Figure 11.1. However, this would require numerous additional locations for the microcell base stations, and we would have to connect all buildings with pico- or femtocells via fiber optics. The densification of the access network could lead to more interference problems, i.e., base stations would interfere with each other more often. Besides, contracts would have to be concluded with the building owners for the pico- and femtocells, but not in the case of WLAN access. The costs for such a network would be relatively high. Moreover, such a network would not correspond to the current philosophy of a mobile network operator, which is to serve all users outdoors and indoors everywhere. However, in addition to the current advantages concerning NIR and 5G, there would also be the great advantage of relying on low-emission network architecture and infrastructure for future developments towards ever higher bit rates and the frequency ranges used for them (see Chapter 12). This could facilitate future developments towards 6G.

11.5 Energy Requirements, Raw Materials, and Sustainability

In the past, the consumption of electrical energy by a mobile network accounted for 15% of operating costs, and now even 20 to 40% [230]. Of these, up to 75% are accounted for by the RAN, i.e., the base stations with their active and passive components [230]. However, base stations are often only used to a low degree over a 24-hour period. Much energy is also required for system signaling via broadcast signals, cooling, etc. [144].

An increase in energy efficiency in the RAN is therefore possible via:

- Measures to reduce power consumption when a base station has no data to send
- A reduction in the energy consumption of auxiliary equipment for air conditioning, power supply, etc.

- Optimization of the efficiency of the hardware, especially when operating significantly below the maximum load [144].

As a result, 5G-NR has reduced the amount of signaling and control in RAN to about 20% compared to 4G-LTE. This gives much more flexibility to put a base station into sleep mode again and again – even for short time periods – during times of low traffic volume, thus saving energy and significantly increasing energy efficiency [144; 101].

A second lever for increasing energy efficiency is seen in the increased use of micro-, pico-, and femtocells, which is already necessary for the medium term due to the frequency spectrum used and the higher data rates of 5G as a result of the lack of traffic capacity. Because of the close distance to the users, advantageously such base stations work with much lower transmission powers. This also applies to the terminal equipment. Besides, there are many fewer UEs in such a small cell. In this respect, the sleep mode can be used more often and longer at certain times, and the base station can even be switched off entirely at certain times (e.g., at night in a shopping mall). However, we should not forget that more base stations also result in more additional devices – at least the power supply – and thus, the corresponding energy consumption increases [144].

In [120], further reasons for better energy efficiency with 5G are mentioned, which are, however, partly related to the already mentioned approaches of sleep mode and small cells:

- Higher data rates and lower latency: For the same data volume, significantly more times of inactivity and, therefore, sleep mode occur compared to 4G.
- The protocols for control and signaling in the RAN compress the messages. This also increases inactivity times. The MPTCP (Multipath TCP) used, if applicable, ensures a very reliable message transport. Messages must be sent repeatedly exceptionally rarely. The energy efficiency increases accordingly.
- The massive MIMO antennas used to offer a higher antenna gain. Beamforming focuses the transmission power and minimizes interference. This leads to lower energy consumption in the base station and terminal equipment.

From the above, one could conclude that 5G has a clear advantage over 4G in energy consumption. This certainly applies to the power consumption per transmitted bit in W/bit. However, with the appropriate offerings and more data-intensive services, the amount of data to be transmitted will increase due to corresponding customer demand, which partially uses up the energy savings. Besides, massive MIMO systems with, e.g., 64T64R (64 Transmit, 64 Receive) of the first generation still had a significantly higher energy consumption than normal systems with passive antennas. However, the vendors all stated that this can be lowered with the next versions. At high cell loads, the energy consumption per transmitted bit with massive MIMO is, in any case, lower than with conventional antenna systems. In addition, even

with 4G, a base station can use several frequency bands simultaneously. This situation is the usual case with 5G because of the only fragmented usable frequency ranges. With each additional frequency band, the power demand increases. According to [108], the result is a 70% higher power consumption for base stations for macrocells, and even significantly higher in the case of more than 10 frequency bands. In this respect, the energy savings achieved by the RAN system optimizations for 5G are offset by a possible increased power requirement due to more massive amounts of data to be transmitted, a relatively large number of frequency bands, and only gradual hardware optimization.

The overall result is that the energy efficiency benefits of 5G described above can only be made usable in practice through careful network and component design. Measures for the RAN for this purpose are summarized in Table 11.2.

Tab. 11.2: Impacting factors and measures for RAN energy consumption optimization in 5G [230; 231; 232]

Impacting factors	Measures for RAN energy consumption optimization
Network architecture	<ul style="list-style-type: none"> – As much central and as little outdoor technology as possible – Simple mobile radio sites without equipment rooms and air-conditioning units – With C-RAN (see Section 3.1) and thus centralized BBUs, reduction of power supply and air-conditioning costs – Use of energy-efficient combinations of frequency bands – Minimization of the RAT technologies used, e.g., only 5G-NR
Energy efficiency of the equipment	<ul style="list-style-type: none"> – Replacement of outdated hardware, e.g., 3G with 5G RAN technology. – Highly integrated base stations (BS) as possible, with multi-band and MIMO technology – Processing of as many frequency bands as possible in one BS module, e.g., 1 module for 700-, 800- and 900-MHz bands, 1 module for 1800-, 2100-, 2600- and 3500-MHz bands – Use of energy-saving functions such as standby and sleep modes – Optimized coordination of power consumption, power supply, power storage, and air conditioning depending on traffic volume
Radio performance and traffic volume	<ul style="list-style-type: none"> – Highest possible interference and spectral efficiency – High antenna gain – Use of multi-antennas, e.g., 64T64R, for high energy efficiency per bit – Software-controlled adjustment of transmit power and set to standby or sleep modes depending on traffic volume, possibly AI-supported

Nevertheless, it is questionable whether the requirement in ITU-R Recommendation M.2083 [128] (see Section 4.3) for RAN energy efficiency of 100 times better than IMT-Advanced (4G), i.e., the same energy consumption at 100 times performance is implemented.

As explained in Section 9.6, a 5G system uses cloud resources in the (R)AN, but especially in the 5G core. Thus, MEC applications and C-RAN functions are provided in the edge cloud, the 5GC functions in a central cloud. This leads to a high and growing demand for local, regional, and central data centers. In [105], a distinction is made between small-scale data centers with a power requirement of up to 5 kW, campus data centers up to 20 kW, edge-cloud data centers with more than 100 kW, and highly scalable data centers with a power requirement of 10 MW or more. Overall, the study [105] identifies increased energy consumption in data centers due to the introduction of 5G for Germany. With a share of 5G IP traffic of almost 14%, electrical energy consumption will rise to around 2.6 TWh by 2025, with a total of about 19 TWh for all data centers together. This means that introducing 5G will increase the energy requirements of data centers in Germany by approx. 16%, and this already by 2025.

Based on this predicted energy consumption situation at 5G with correspondingly high CO₂ emissions, we have to consider counteractive measures. These are, first, the measures for the RAN discussed above, and second, the use of renewable energy for the power supply of the 5G systems in operation and production. Besides, old, energy-inefficient 2G and 3G technology should be switched off as soon as possible if it has not already been done [145].

As already mentioned, in the RAN and 5G core, energy consumption in the utilization phase of the system components dominates the environmental balance. The environmental impact of mobile devices and sensors, on the other hand, is determined by the raw materials required and the manufacturing processes. In this respect, the energy- and raw material-intensive semiconductor and printed circuit board production have a high environmental impact. With increasingly higher frequencies, multiple antennas and beamforming, even more powerful semiconductors are required. This also changes the material mix. In addition to standard silicon oxide, gallium arsenide, gallium nitride, or silicon-germanium are increasingly being used. The environmental impact of these types of components has not yet been thoroughly researched. However, it is a fact that more attention needs to be paid to environmental assessment and ecodesign to implement sustainable communication systems in the long term [167].

Finally, and in a summary of this section, the sustainability or the contribution to the sustainability of 5G networks should be discussed. In 1987, the so-called Brundtland Commission, the UN World Commission on Environment and Development, published the report "Our Common Future", in which the concept of sustainable development was formulated and defined for the first time: "Sustainable development meets the needs of the present without compromising the ability of future

generations to meet their own needs." [233] Sustainable development is thus development that meets the needs of the present generation without compromising the ability of future generations to meet their own needs and choose their lifestyles.

Based on this, among other things, the UN General Assembly introduced 17 Sustainable Development Goals (SDGs) in 2015 in the agenda "Transforming our world: the 2030 Agenda for Sustainable Development" [234]. The guiding principle of the 2030 Agenda is to enable people worldwide to live with human dignity while protecting the natural basis of life in the long term. This includes economic, ecological, and social aspects.

In [229], the 17 UN SDGs are discussed concerning the possible contributions of mobile networks and ICT. It is necessary to consider that the mobile communications technology developed and implemented should be sustainable and used sustainably. The former includes aspects such as low energy consumption, use of non-toxic materials, environmentally friendly network equipment, including its placement in locations that do not negatively impact the environment, and sustainable supply chains. The second area, the sustainable use of mobile, can be seen, for example, in the provision of digital services in developing countries for micro banking, microfinance, enabling, and access to markets. In general, universal access to information is a critical factor in achieving the United Nations SDGs. Concerning achieving sustainability goals, however, it is crucial to consider that there may be adverse interactions or that trade-offs may have to be made. For example, placing network equipment in a remote area may improve the quality of life and job opportunities but have negative impacts on wildlife there.

Table 11.3 shows, according to [229], the UN SDGs and possible contributions of mobile networks, supplemented by references to specific related contributions by 5G technology.

Tab. 11.3: UN Sustainable Development Goals (SDGs) and possible contributions of mobile networks

SDGs [229]	Existing linkage with mobile communications/ICT [229]	Special 5G contribution
1 No poverty	Communications infrastructure to stimulate local economic growth in poor communities. Lowering barriers to economic resources for people in extreme poverty via access to mobile money, microfinance, and the creation of employment opportunities	
2 Zero hunger	Can help farmers increase crop yields and business productivity through better access to market information, weather forecasts, training programs, and tailored information	mMTC, IoT, sector-specific solutions (verticals)
3 Good health and well-being	Enables communication with healthcare professionals, health monitoring, access to healthcare programs,	mMTC, IoT, eMBB, security

SDGs [229]	Existing linkage with mobile communications/ICT [229]	Special 5G contribution
	healthcare identity management, and big data provision during epidemics. It also enables access to digital medical records and AR/VR for medical education.	
4 Quality education	Promotes digital learning and enables access to learning content anytime, anywhere. Can be used in literacy, numeracy, and tutoring. Mobile learning can help overcome economic barriers, urban-rural and gender gaps	eMBB, satellite communication
5 Gender equality	Mobile devices allow women in low-income countries to feel safer, be better connected, and access information, services, and life-enhancing opportunities. Participation in the sharing economy is enabled. Women are better able to make their voices heard in their communities, with their governments, and at the global level	eMBB, security, satellite communication
6 Clean water and sanitation	Enables smart water management, facilitates measurement, monitoring of water supply, and necessary actions. Can help distribute water equitably and sustainably on-site, ensure sanitation and hygiene, and optimize operations	mMTC, IoT, verticals
7 Affordable and clean energy	Supports data analytics for renewable energy sources, green energy procurement, smart grids, distributed energy systems, and smart metering	mMTC, IoT, URLLC, energy efficiency
8 Decent work and economic growth	Expands the market through online channels to consumers and provides access to mobile financial services. Changes the way business is done and creates new employment opportunities	Verticals
9 Industry, innovation and infrastructure	Provides the infrastructure for low-cost voice and data services and offers employment opportunities even in remote areas. As a broadband infrastructure, it is a key driver of industry and innovation.	Network slices, MEC, campus networks, satellite communication, eMBB, verticals
10 Reduced inequalities	Enables access to information and social networks, promotes social and political participation, provides access to marketplaces and mobile money, and digital identity management. Enables disadvantaged segments of society to access information and knowledge	eMBB, security, satellite communication
11 Sustainable cities and communities	Teleworking also makes rural areas attractive for employees. Collaboration systems enable collaboration between companies and research institutions, and increase business opportunities in rural areas. Using drones can boost productivity of rural business, improve access to goods, and reduce production and delivery costs.	mMTC, IoT, URLLC, satellite communication

SDGs [229]	Existing linkage with mobile communications/ICT [229]	Special 5G contribution
12 Responsible consumption and production	Cloud computing, smart grids, and smart metering can reduce energy consumption. ICT energy consumption and negative impacts such as e-waste need to be minimized.	MEC, URLLC, mMTC, IoT
13 Climate action	Helps other sectors avoid greenhouse gas emissions, enables smart traffic management, urban lighting, parking, logistics, energy management systems, remote working, sharing economy, smart grids, connected health, and precision agriculture	mMTC, IoT, URLLC, verticals, energy efficiency
14 Life below water	Facilitate conservation and sustainable use of the oceans through improved monitoring and reporting, with timely and accurate data, with sensors providing real-time updates. Provide the necessary communication links for autonomous ships and underwater vehicles	mMTC, IoT, satellite communication
15 Life on land	Supports conservation and sustainable use of ecosystems and the prevention of biodiversity loss through improved monitoring, reporting, and provision of timely and accurate data on a global basis and, via sensors, on a local basis. Big data can be used to analyze short- and long-term trends and plan mitigation measures for biodiversity, pollution, weather, and ecosystems.	mMTC, IoT, MEC
16 Peace, justice and strong institutions	Supports agencies, such as police, in preventing violence and adheres to strict data privacy and security policies in accordance with national and international law. Promotes the use of open data, increases transparency, empowers citizens, and promotes economic growth by recording and tracking government and local demographic data. Enables communication and data transfer in disasters	Emergency services, security
17 Partnerships for the goals	Supports collaboration between the public and private sectors. Acts as a catalyst for the 3 pillars of sustainable development: economic growth, social inclusion, and environmental sustainability	eMBB, URLLC, mMTC, IoT, verticals, campus networks, energy efficiency

Essential contributions to a sustainable 5G system are the above-mentioned energy efficiency measures, manufacturing, and operation with renewable energy, the use of recyclable materials, and, in general, an environmentally friendly overall life cycle from material selection through development, production, transport, and operation to disposal. This is supplemented by a recycling-oriented product life cycle with environmentally friendly materials, waste-minimizing product develop-

ment, and optimized packaging. Overall, the dependency on natural resources and the CO₂ emissions caused by 5G must be minimized [230; 231; 232].

However, the impact of 5G in terms of sustainable use could be significantly greater than the sustainable development, production, and operation of 5G systems. This is because 5G was designed to support a wide range of industries (verticals) in the best possible way and can have a significant impact on their energy efficiency. Examples include 5G usage for smart energy management systems in buildings and public infrastructure, smart metering and smart grid, IoT-based monitoring, and AI-powered analytics for a predictive shutdown of devices in urban infrastructure systems or private households. Improved conferencing capabilities and VR applications can save office space and travel. Passenger and freight traffic can be optimized. 5G, in tandem with cloud computing, AI, and IoT, is a catalyst for significant increases in energy efficiency, particularly in the healthcare, manufacturing, transportation, and energy supply industries. Smart cities and intelligent transport systems, together with 5G, could potentially improve energy efficiency in many areas of life [232].

12 Future Developments

As stated in Section 5.2 on standardization, 3GPP Release 16 for 5G has been fully standardized by the end of 2020. A 5G system compliant with Release 16 corresponds to the IMT-2020 target system defined by the ITU (see Sections 4.3, 5.2, and 6.2). The next step in further developing 5G was Release 17, as already mentioned in Section 4.3 and Section 5.2 regarding standardization. But this is certainly not the end of the 5G evolution. Release 18 is already in progress since 2021. Standardization of Release 19 was started in 2022. 3GPP Release 18 is, therefore, discussed in more detail in Section 12.1, and important new topics are mentioned for Release 19. Besides, people are already thinking about the networks after the 5G era. The activities of the ITU-T under the keyword Network 2030 are essential for this. These are examined in Section 12.2. 6G is still in the early stages of standardization but is, of course, a major topic in research. For related results, please refer to Sections 12.3 to 12.6.

12.1 Further Development of 5G

As already mentioned, the next steps in further developing 5G is the 3GPP Release 18, also called 5G Advanced, and implementations of 5G systems based on it. The most crucial service requirements, which complement Releases 15, 16, and 17, are summarized in Table 12.1 [238]. Particularly worth mentioning here are

- Ranging-based services
- Personal IoT Networks (PINs) and Customer Premises Networks (CPNs)
- Support of AI/ML models
- Mobile base station relays
- Tactile and multi-modal communication services.

Tab. 12.1: Additional service requirements for 5G system in 3GPP Release 18 [238]

Service Requirements

Basic capabilities

Extensions to network slicing, e.g., regarding UE capabilities, for multiple access of a UE, for UE power consumption, for roaming situations

Support for multiple terrestrial and/or satellite-based wireless backhaul links

Extensions to expose network functions to 3rd party providers, e.g., for authentication and authorization and information on RAN capabilities

Extensions to QoS monitoring

Extensions to 5G LAN, e.g., for 3rd party

Communication services, especially for Smart Grid

Service Requirements

Extensions for service function chaining, e.g., per single UE, for 3rd party, for roaming
 Improvement of the timing resiliency
 Ranging-based services, e.g., relating to the distance between 2 UEs, also in the 3D use case
 Personal IoT Networks (PINs), e.g., with wearables, and Customer Premises Networks (CPNs)
 Extensions to IMS-based multimedia services, e.g., for AR support, 3rd party providers
 Support of AI/ML models incl. splitting of AI/ML operations, e.g., between server in network and UE
 Access to local services, also temporary, via hosting networks (public or non-public), e.g., at an exhibition center or in a shopping mall
 Mobile base station relays
 Tactile and multi-modal communication services that enable applications with more than one input source and/or output destination, including synchronization, e.g., with modalities audio, video, via sensors for brightness, temperature, humidity, etc., haptic data on pressure, texture, vibration, etc.

Performance requirements

For video surveillance via satellite link and up to 120 km/h speed of the UE, data rates of 0.5 Mbit/s DL and 3 Mbit/s UL
 For timing resiliency, among other things, up to 250 ns synchronization accuracy
 For ranging-based services, among others, up to 10 cm accuracy at a distance of 30 m
 For AI/ML model transfer, e.g., 2 ms E2E delay at 1 Gbit/s and 99.999 % availability
 For tactile and multi-modal communication services, for multi-modal navigation, e.g., 50 ms E2E latency, up to 2 Mbit/s data rate, and 99.999 % reliability with haptic feedback

By mid-2023, the key features and functionalities beyond the scope of Release 17 have been defined. Table 12.2 provides an overview that has not yet been finalized. Particularly noticeable here are the new topics that already point to 6G (see Sections 12.3 to 12.6):

- XR (eXtended Reality)
- AI/ML (Artificial Intelligence/Machine Learning)
- Personal IoT Networks.

Tab. 12.2: Features and functionalities of 3GPP Release 18 [58]

Features

New

XR services
 Support for AI/ML-based services
 Satellite-based backhaul connectivity
 Resilient timing
 Support of Personal IoT Networks

Features

Vehicle mounted relays
Ranging-based services
Support of service function chaining
Media capabilities for AR glasses
XR conversational services
WebRTC-based services and collaboration
AI/ML-based management
Ad hoc groups for critical communications, e.g., in the event of a disaster
Exposure of network capabilities for IoT platforms
NR network-controlled repeaters
NR support for UAV
Mobile Terminated-Small Data Transmission (MTS DT) for NR
AI/ML for NG-RAN

Improvements and enhancements

Edge computing
Network automation
Non-public networks
Network slicing
Location-based services
Multicast/broadcast services
Satellite access
IMS-based multimedia services
ATSSS

Exposure of services and network functions for 3rd party providers

UAS, UAV, and UAM support (Urban Air Mobility)

Terminal devices with reduced capability (RedCap)

Audio and video codecs

5G Media Streaming (5GMS)

RAN self-configuration

Network data analytics

Energy efficiency

V2X

MIMO

Network coverage

Dynamic spectrum sharing

NR and IoT NTN

QoE

FR1 (n105: FDD, 663 - 703 MHz UL, 612 - 652 MHz DL [239]) and FR2 frequency ranges

Features

NR for high-speed trains with FR2

Figure 12.1 shows the timetable for the standardization of Release 18 for 3GPP, with the standards being finalized by summer 2024. At the end of the standardization work, the specified features are documented in TR 21.918.

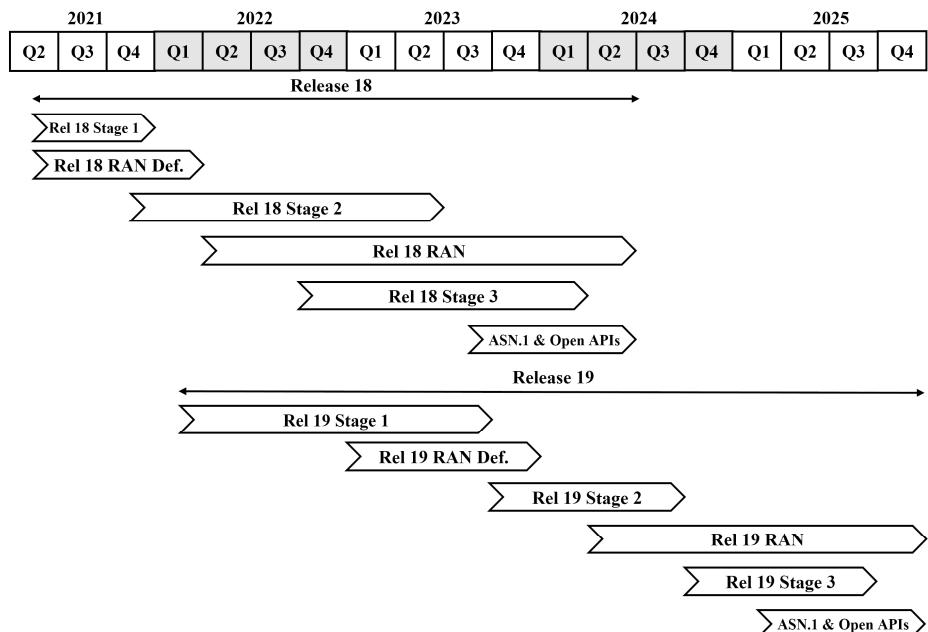


Fig. 12.1: Time schedule for the standardization of 3GPP Release 18 [58; 237; 57]

Figure 12.1 also indicates the subsequent Release 19. Its standardization has already begun in 2022. Compared to the previous releases, new topics are addressed [268]:

- Network of service robots with ambient intelligence
- Energy efficiency as a service criterion
- Roaming of value-added services
- Integrated sensing and communication
- IoT with devices that harvest their power from the environment (e.g., from radio waves or light)
- Localized mobile metaverse services.
- Network sharing aspects.

The trend toward 6G also becomes apparent (see Sections 12.3 to 12.6).

12.2 Network 2030

As mentioned above, the ITU-T Study Group 13 has established a Focus Group on Technologies for Network 2030 (FG NET-2030) in July 2018 with the goal “Network 2030: A pointer to the new horizon for the future digital society and networks in the year 2030 and after that.” FG NET-2030 aims to determine the fundamental properties of networks in 2030 and beyond. This is based on the assumption that new application scenarios must be supported by holography, with extremely fast response in critical situations and with high-precision localization. In this context, questions regarding future network architecture and the required functionalities and mechanisms have to be answered [99].

The FG NET-2030 was organized in three subgroups:

- Use Cases & Requirements (Sub-G1)
- Network Services & Technology (Sub-G2)
- Architecture & Infrastructure (Sub-G3).

The first results were available in a comprehensive white paper [123] and a first sub-G2 results paper [168].

The next steps in multimedia applications after AR and VR could be holography and communication with more senses, i.e., not only by seeing and hearing but also by feeling (passive, tactile perception) or touching (active, haptic perception), smelling and tasting [123]. With holography, three-dimensional objects can be detected and subsequently projected in free space by exploiting the wave character of light with coherence and interference. These objects can be people, things, or objects in general. If sequences are scanned, sequences or changes can also be visualized. A three-dimensional view of the hologram on site is possible without 3D glasses.

In the context of the FG NET-2030 considerations, such holographic data should not only be recorded locally but transmitted or streamed via the network. One speaks then of Holographic-Type Communications (HTC). This enables novel applications such as

- holographic telepresence of people (digital twins) in a room for a meeting,
- transmission of holographic representation of an object (difficult to access), e.g., a machine to be repaired,
- telesurgery or also
- practical training across distances [123].

Such holography applications lead to enormous demands on future networks. In addition to the transmission of the resolution, color depth, and image sequence required for video, spatial data for different viewing angles depending on the position of the viewer, as well as synchronization data, must also be transmitted. This leads to extremely high bit rates for high-quality holograms in real-time.

Starting from the point of view of the hologram, “impressions of the hologram or avatar” recorded by cameras and microphones could be transmitted in the reverse direction. This means that the HTC would be combined with video and audio streams, which requires correspondingly high-precision synchronization.

Another extension could be the combination of HTC with feeling or touching. The “touching” of a hologram could be transmitted back. A possible application would be the repair of a real machine over distances or a remote surgery by appropriate measures on the corresponding hologram. Such holographic applications require extremely short delay times due to the fast reactions to the “touches” expected by the user and, in the case of remote surgery, extremely high availability.

As already mentioned above, other senses, such as taste and smell, should also be included. These are caused by (chemical) reactions of the corresponding active ingredients, which are perceived by a human being with his corresponding receptors. Therefore, the main challenge here is to make corresponding actuators available. An example is a so-called digital lollipop, an electronic device that can synthetically create a taste by stimulating the human tongue with electric currents [123].

FG NET-2030 also sees further requirements for future networks in industrial automation, autonomous systems, e.g., in traffic as well as in large sensor networks. Among other things, increased requirements on latency are seen here: time delays below 1 ms in industrial control loops or the defined, mandatory timely arrival of messages in traffic control despite thousands of simultaneously communicating vehicles, traffic lights, etc. in a comparatively small space. The latter is not only a question of latency but also of the possibility of the exact synchronization of many participating systems. High-precision synchronization is also important for applications such as online gaming or collaboration with many participants at different locations. All in all, the network must provide a synchronized view of a specific application with a wide variety of geographically distant information sources and sinks [123].

Besides, future networks should offer much more extensive possibilities for emergency situations, including earthquakes and floods. Precise locations must be available immediately, and optimal navigation must be ensured, etc. In this context, HTC, AR, and VR, as well as tactile applications, should also be used [123].

[240; 241] summarizes a number of possible Network 2030 use cases and makes relative assumptions about the resulting requirements. For this, we can categorize the requirements as follows:

- Bandwidth: bandwidth, capacity, QoE, QoS, flexibility, and adaptable transport
- Latency: latency, synchronization, jitter, accuracy, scheduling, coordination, and geolocation accuracy
- Security: security, privacy, reliability, trustworthiness, resilience, traceability, and lawful intercept
- AI: data computation, storage, modeling, collection and analytics, autonomy, and programmability

- ManyNets (coexistence of heterogeneous networks): addressing, mobility, network interface, and heterogeneous network convergence.

Based on the above requirements, the five requirement categories of bandwidth, latency, security, AI, and ManyNets for twelve use cases of a Network 2030 were relatively evaluated with scores ranging from 1 (not important) to 10 (extremely important) according to Table 12.3 [240; 241].

Tab. 12.3: Network 2030 use cases and evaluated requirements [240; 241]

Use case	Requirements				
	Bandwidth	Time	Security	AI	ManyNets
Holographic type communications (HTC)	10	7	5	5	1
Tactile Internet for remote operations (TIRO)	5	10	7	3	2
Intelligent operation network (ION): Intelligent (AI) monitoring and control	3	5	8	10	4
Network and computing convergence (NCC): Computing-aware network capabilities	5	10	5	5	3
Digital twins (DT)	6	8	8	6	5
Space-terrestrial integrated network (STIN)	5	7	7	2	10
Industrial IoT (IIoT) with cloudification	8	10	8	3	8
Huge Scientific Data applications (HSD): e.g., for astronomical telescopes, particle accelerators	10	7	5	5	1
Application-aware data Burst Forwarding (ABF): e.g., for video surveillance with real-time image processing	5	5	7	3	2
Emergency and disaster rescue (EDR)	3	5	8	10	4
Socialized Internet of Things (SiIoT): a decentralized approach to foster the interactions among trillions of objects	5	9	5	5	3
Connectivity and sharing of pervasively distributed AI data, models and knowledge (CSAI)	6	8	9	6	5

According to Figure 12.2, Network 2030 is intended to provide a solution for the fully networked digital society that integrates new industries (verticals), enables innovation, e.g., through holography, and offers new communication services with extreme requirements via interconnected new network infrastructures.

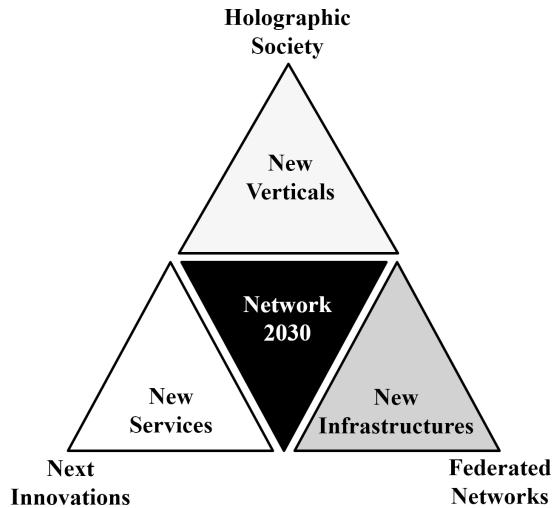


Fig. 12.2: Network 2030 vision [123]

Figure 12.3 shows the most important industries and their applications and the resulting demand for a network in 2030 [123].

Social and Entertainment 2030	Healthcare 2030	Automotive 2030	Education 2030	Factories 2030
Telepresence (real room experience with video conferencing)	Tele surgery	Coordinated (timed)	Coordinated	Autonomous
	Tele monitor (patient monitoring)			
Holoportation (transfer of the field of view as 3D model)	Tactile	Situation response (reaction to critical situation)	In-room presence	Automation
Multi-sense	Time awareness (real-time)	Time awareness	Holographic media	Time awareness
Holographic media	Haptics	Tactile	Haptics	Tactile

Fig. 12.3: Most important industries and their applications for a Network 2030 [123]

Network 2030 refers to an integrated, highly automated, intelligent infrastructure containing a number of operator domains in various types of network segments (e.g., wired/wireless access, core, edge, and space segments). This integration is based on a dynamic interaction between computing, storage, and network services/applications resources/devices in all network segments.

Network 2030 is envisaged to support different and very stringent functional and non-functional requirements, including the strict low latency and the large volume of data exchange requirements. In some cases, these requirements are to be supported per network slice basis. Additional Network 2030 new composite characteristics and capabilities are [242]:

- Enhancing IP best effort service provision with service quality information, network conditions enablers to achieve guarantees for KPIs or QoS as required by future precision services and applications per slice.
- Evolution towards native support network functions for very low latency, very high bandwidth, very high reliability/resilience, trustworthiness, and privacy, delivering stringent non-functional requirements with guarantees for KPIs/QoS per slice needed for future network service
- Determinism in delays and lossless transmission
- Native support for multiple types of delivery services, in-time/on-time service activation, and availability
- Elasticity in network services customization and network functions componentization
- Effective programmable network protocol and flexible dynamic transmission
- Intrinsic secured networking and trust networking
- Higher levels of robustness in the face of failures
- Integration of large numbers on intelligent methods (AI/ML (Machine Learning) based methods) in the network infrastructure, control, and management
- Evolution towards intent-driven distributed management of all physical and virtual network elements and network functions.

In terms of performance, [168] requires end-to-end delays down to less than 1 ms (tactile applications), a packet loss rate close to 0, bit rates of 1 Tbps and above (holography applications), and 1 ms synchronization accuracy (tactile applications).

In [124], based on the results for a Network 2030, the application scenarios outlined below are derived according to Figure 12.4:

- Very Large Volume & Tiny Instant Communications (VLV&TIC),
- Beyond Best Effort & High Precision Communications (BBE & HPC)
- ManyNets.

For comparison, please refer to the ITU-R IMT-2020 application scenarios (5G) eMBB, URLLC, and mMTC in Section 4.1 and especially in Figure 4.1.

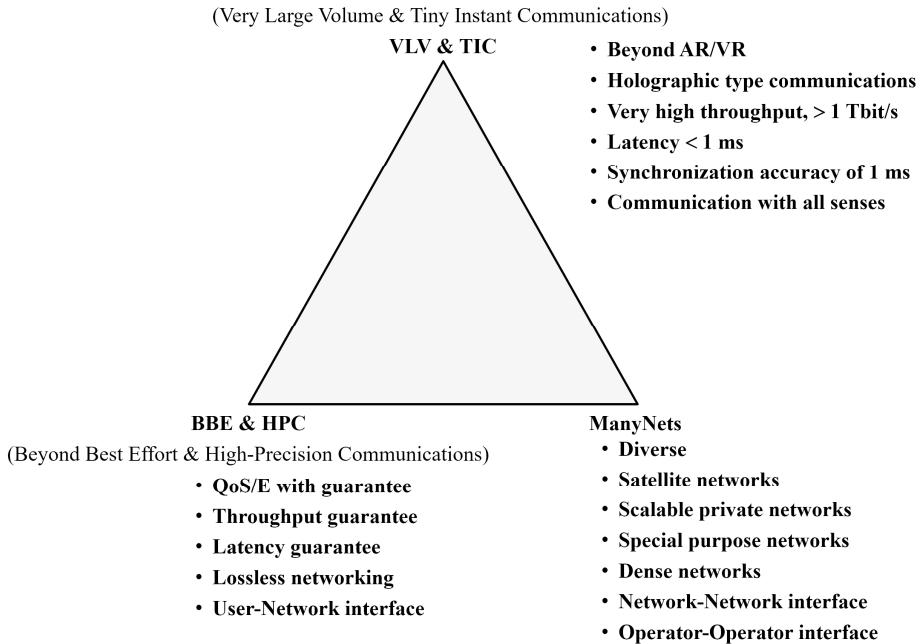


Fig. 12.4: Application scenarios for a Network 2030 according to ITU-T FG NET-2030 [124]

In a first approach [242] describes the required network architecture and formulates nine architectural principles for it:

- Simplicity: With the proliferation of virtualization, networks will consist of many virtualized and non-virtualized components, which makes the Network 2030 complex. Complex systems are generally less reliable and less flexible. The complexity of an architecture is proportional to its number of components. One way to increase the reliability or flexibility would be to reduce the number of components in a service delivery path (i.e., a service chain or a protocol path or a software/virtual path).
- Native programmability: Network functions should be able to be composed in an “on-demand”, “on-the-fly” basis. Programmability in networks refers to executable code injected into the execution environments of network elements to create the new functionality at runtime. Different services can effectively call any network function component and/or resources on-demand flexibly and quickly, based on the automatic allocation and elastic capacity expansion of the underlying network resources.
- Backward compatibility: E.g., the network needs to be capable of supporting, unifying, and integrating protocols supporting various services that optimally meet the needs of new (e.g., IoT devices) but also existing devices.

- Heterogeneity in communication, computing, storage, service, and their integration
- Native slicing
- Unambiguous naming of network functions and services: E.g., user systems are not accessing a specific server anymore, but the content, function, or service that the server would host.
- Intrinsic anonymity and security support for all network operations
- Resilience
- Network determinism: For fulfilling end-to-end requirements of new business applications such as industrial control, telemedicine, robotics, and vehicle networking, the network needs to introduce explicit determinism in very stringent non-functional requirements (accessibility, availability, certification, consistency, compliance, extensibility, fault tolerance, integrability, interoperability, maintainability, operability, performance, privacy, resilience, reliability, robustness, scalability, security) with guarantees per partitions of the infrastructures.

Figure 12.5 shows the relations between these principles, the requirements, and architecture.

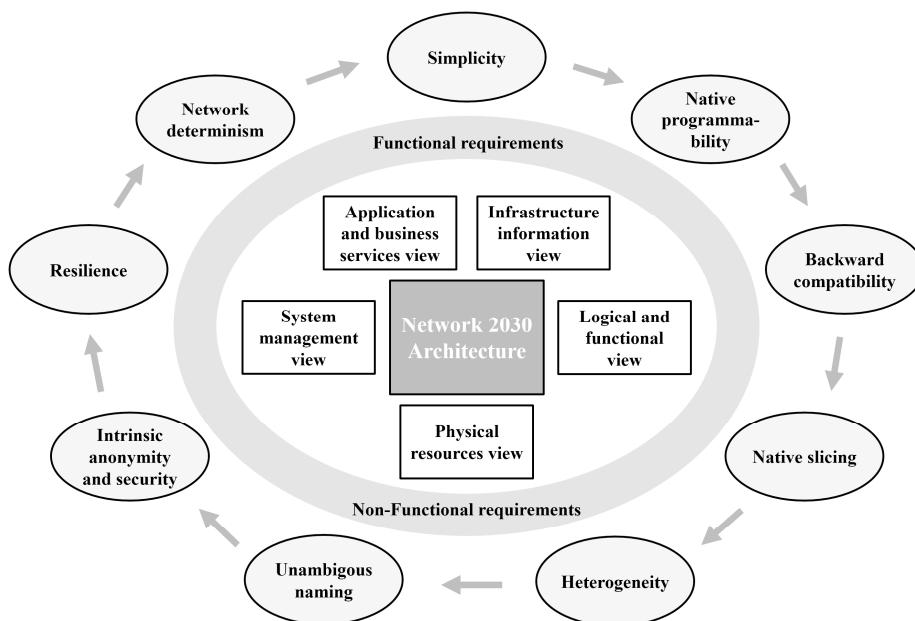


Fig. 12.5: Network 2030 architecture principles [242]

[168] presented the first concrete considerations for technical implementation. Among other things, they consider a new packaging procedure in which the user data of a packet are divided into chunks according to their importance and priority for the provision of the service. Metadata makes this information available. The packet source defines it. When forwarding a packet, a router will not discard the whole packet in an extreme case, but only the less relevant parts, e.g., only the P and B frames, not the I frames of an MPEG4 video stream (Moving Picture Experts Group). The receiver must then be able to further process the modified packet concerning the service provision, e.g., to display the video only with I-frames.

In addition, [242] contains considerations about the different functional areas of a Network 2030, such as access network and edge, space networking, routing and addressing, security, privacy, and trust, QoS, burst switching, network slicing, network management, quantum computing.

Of the above-mentioned functional areas of a future Network 2030, we would like to look at the application-aware Burst Switching. A burst is the basic data unit that can be processed by the application. For example, a burst can represent a photo in an image processing system or a video clip in a video streaming service. The burst forwarding network uses burst as the basic transmission unit. The data source sends the entire burst using the line rate of the network interface card (NIC). End-to-end virtual channels are created for the burst transmission. Burst switching is beneficial for applications where large blocks of data from different sources, e.g., video cameras, must be processed centrally within a defined time, e.g., in the cloud. The accumulated bit rate of the sporadically transmitting sources is much higher than the bit rate for accessing the cloud. Burst switching now ensures that always complete bursts with the maximum bit rate required by the source are transmitted through the entire network to the sink. This leads to increased throughput due to congestion-free transmission of the bursts and an increase in processing efficiency. For forwarding the bursts through the network, a burst is divided into smaller packets, so-called burstlets (packets), and transmitted interleaved with the burstlets of other bursts (up to the maximum end-to-end bit rate). Bursts that are too many at this time are buffered at the input of the network [242].

12.3 Research, Regulation, and Standardization on 6G

Apart from the activities of the FG NET-2030 (see Section 12.2), which relate to a general network or Internet for the time after 2030, the considerations on 6G are all still at the research stage. The example of early and pooled research on 6G represents the 6G Flagship project [67] of the Finnish University of Oulu. Initiated by this project, the first 6G Wireless Summit conference was held in 2019 [68]. Meanwhile, intensive research efforts on 6G, primarily organized in alliances and with public funding, are taking place worldwide. In addition, in early 2021, the ITU-R started its

work on 6G under the heading "IMT towards 2030 and beyond (IMT-2030)" [244; 269]. Before the research and pre-standardization work on 6G is discussed in more detail below, the question of why a 6G network might be needed should first be addressed.

The 5G Infrastructure Association (5G-IA), which represents the private sector as part of European mobile communications research funding, commented in [245]:

- The convergence of physical, human, and digital worlds in 6G requires support for digital twinning (providing digital twins of real systems and processes), immersive (perceiving the virtual environment as real) communication, cognition, and connected intelligence.
- To provide flexibility, programmability should be at the heart of 6G.
- 6G needs to support deterministic end-to-end services.
- 6G needs to provide integrated sensing and communications, enabling high-accuracy localization and high-resolution sensing services.
- 6G plays an ambitious role towards sustainability, to reduce its footprint on energy, resources, and emissions and improve sustainability in other parts of society and industry.
- 6G needs to become a truly trustworthy infrastructure that becomes the basis of societies of the future.
- To ensure that 6G can be inclusive for all people across the world, it needs to be scalable and affordable.
- 6G needs, where necessary, to significantly stretch the KPIs that 5G can achieve now.

According to ITU-R WP 5D, the motivation for the development of IMT-2030 or 6G is „to continue to build an inclusive information society and to support the UN's sustainable development goals (SDGs). To this end, IMT-2030 is expected to be an important enabler for achieving the following goals, among others:

- Inclusivity: bridging digital divides, to the maximum extent feasible, by ensuring access to meaningful connectivity to everyone
- Ubiquitous connectivity: To connect unconnected, IMT-2030 is expected to include affordable connectivity and, at minimum, basic broadband services with extended coverage, including sparsely populated areas.
- Sustainability: Sustainability refers to the principle of ensuring that today's actions do not limit the range of economic, social, and environmental options to future generations. IMT-2030 is envisaged to be built on energy efficiency, low power consumption technologies, reducing greenhouse gas emissions, and use of resources under the circular economy model, in order to address climate change and contribute towards the achievement of current and future sustainable development goals.
- Innovation: fostering innovation with technologies that facilitate connectivity, productivity, and the efficient management of resources. These technological

advances will improve user experience and positively transform economies and lives everywhere.

- Enhanced security, privacy, and resilience: The future IMT system is expected to be secure and privacy-preserving by design. It is expected to have the ability to continue operating during and quickly recover from a disruptive event, whether natural or man-made. Making security, privacy, and resilience key considerations in the design, deployment, and operation of IMT-2030 systems is fundamental to achieving broader societal and economic goals.
- Standardization and interoperability: To achieve wide industry support for IMT 2030, future IMT systems are expected to be designed from the start to use transparently and member-inclusively standardized and interoperable interfaces, ensuring that different parts of the network, whether from the same or different vendors, work together as a fully functional system.
- Interworking: IMT-2030 is expected to support service continuity and provide flexibility to users via close interworking with non-terrestrial network implementations, existing IMT systems and other non-IMT access systems. IMT-2030 is also expected to support smooth migration from existing IMT systems, where including support of connectivity to IMT-2020 and potentially IMT-Advanced devices will be advantageous for inclusivity.“ [146]

In [246], the European 6G Flagship Research Project Hexa-X lists seven reasons for a new network architecture, arguing here from a technical perspective in comparison with 5G:

- Enabling AI: future comprehensive AI deployment for network operations and as a service for user applications (AI as a Service, AlaaS)
- Programmability
- Architecture for a network of networks: comprehensive connectivity and global coverage through flexible integration of a wide variety of networks, including non-terrestrial subnets, sub-THz, and visible light communication (VLC) base stations, device-to-device communication, mesh, ad-hoc, campus networks
- New protocols for new 6G spectrum
- Cloud softwarization and service-based architecture: also for network management and orchestration functions
- Continuum orchestration: down to the end devices (including smartphones, wearables, sensors)
- Consideration of sustainability and regulations.

However, as with 5G (see Chapter 4), 6G is also based on use cases considered important for society starting in 2030. This is discussed in detail in Section 12.4.

As already mentioned, numerous research and development activities towards 6G globally exist. This occurs primarily in international and national initiatives, research projects, forums, and activities preparing standardization and regulation.

In Europe, initiatives and organizations in the context of the EU should be mentioned first:

- 5G PPP (5G infrastructure-Public Private Partnership) [66], a joint initiative of the European Commission and the European ICT industry, initially launched for 5G research funding
- Subsequently, comparable to the 5G-PPP, the SNS JU (Smart Networks and Services Joint Undertaking) initiative [247] was founded to promote 5G and especially 6G research projects in the EU. The 5G-IA, now renamed 6G-IA (6G-Industry Association), represents the private side in the 5G-PPP and the SNS JU initiative. It operates under the name 6G SNS IA (6G Smart Networks and Services Industry Association) [248].

The research project most shaping the European view of 6G is the 6G Flagship Research Project Hexa-X [249], funded under 5G-PPP and running from January 2021 to June 2023. It is continued, funded by SNS JU, as project Hexa-X-II [250] with a duration from January 2023 to June 2025. While in Hexa-X the research focus was on a 6G vision, the basic concepts and possible technologies, the work in Hexa-X-II covers system analysis up to early validation and proof of concept.

The first 6G research program, "6G Flagship" [67], was already launched in Finland in May 2018 and lasts until 2026 under the leadership of the University of Oulu. In this context, the United Nations Sustainable Development Goals (cf. Section 11.5) are pursued for 6G in addition to the research focus areas of wireless connectivity, devices and circuit technology, distributed intelligence, and human-centric wireless services.

In August 2021, four 6G research hubs were formed in Germany to promote collaboration between academia and industry: 6GEM (6G research hub for open, efficient, and secure mobile communications systems) [251], 6G-life (Digital transformation and sovereignty of future communication networks) [252], 6G-RIC (6G Research and Innovation Cluster) [253], and Open6Hub (6G for Society and Sustainability) [254]. They are funded by the BMBF (German Federal Ministry of Education and Research, Bundesministerium für Bildung und Forschung). Overarching these and other 6G R&D projects, the 6G Platform [255] was created to provide scientific and organizational support for the processes required to implement the German 6G program successfully. To achieve this, harmonization with international regulation and standardization is promoted, and opportunities for participation by society and industry are established.

In the UK, the University of Surrey formed an R&D hub on 6G with the 5G/6G Innovation Centre [263].

In the USA, numerous universities and research institutions are engaged in research on 6G [264]. The private-sector Next G Alliance [256] is very active. It has set itself the goal of massively strengthening mobile communications technology over

the next few years throughout the entire life cycle, from research and development to standardization, production, and market rollout.

In Asia, there are several country-specific initiatives. Industry, science, and government in Japan formed the Beyond 5G Promotion Consortium (B5GPC) [257]. In South Korea, Samsung is very active concerning 6G-R&D. In China, the IMT-2030 (6G) Promotion Group [258] represents the leading platform for promoting 6G R&D, combining the forces of Chinese industry, universities, and research institutions. As a company, Huawei, in particular, should be mentioned in this context. From India, contributions to IMT-2030 come from the TSDSI (Telecommunications Standards Development Society, India) [265].

Without direct reference to countries or government organizations, two further initiatives should be mentioned concerning 6G. One is the NGMN Alliance (Next Generation Mobile Networks) [143], an international alliance of major mobile operators and providers, manufacturers, and research institutes to develop requirements and support standardization for 5G and 6G from the mobile operators' perspective. The other is the one6G Association [259], a global, nonprofit, noncontributory interest group comprising manufacturers, universities, research institutes, network operators, and users.

At the beginning of 2021, the ITU-R Working Party 5D (WP 5D) - IMT Systems (International Mobile Telecommunications) began initial work on the standardization and regulation of "IMT for 2030 and beyond" [244; 269], i.e., for a 6G system. As a first result, there is already a report on the corresponding technology trends [260] and a draft for a "framework and overall objectives of the future development of IMT for 2030 and beyond" [146]. The entire schedule with requirements dated for early 2026, technology proposals by early 2029, and IMT-2030 specifications by mid-2030 are shown in Figure 12.6 [261; 269].

Figure 12.6 already considers that frequency allocations are also required for an IMT-2030 or 6G system. Frequency allocations are made in global coordination every three to four years at the ITU World Radiocommunication Conferences (WRC). For 6G, this occurs at the WRC-23 conference at the end of 2023 [262] and WRC-2027.

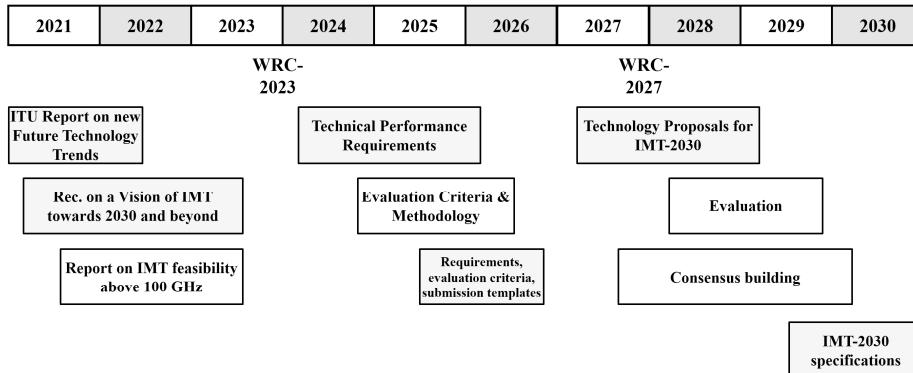


Fig. 12.6: ITU-R timetable for work on "IMT towards 2030 and beyond" [261; 269]

Regarding the frequency ranges for 6G, [266] assumes the following considerations. There will be significant improvements in the eMBB, URLLC, and mMTC deployment scenarios known from 5G. Due to more powerful communication techniques and technological advances, e.g., in sensors, imaging, displays, and AI, entirely new applications become possible. These include immersive extended reality (XR), mobile true-to-life holograms, digital twins, etc. This may require data rates of up to several Tbit/s. This, in turn, requires bandwidths of several hundred MHz up to several GHz. This makes it obvious that new frequency ranges must be opened up compared with 5G. On the one hand, this concerns the mid-band spectrum, under which the range from 1 to 7 GHz has been counted up to now. It should be extended to 24 GHz for 6G to gain a few hundred MHz of contiguous bandwidth and thus achieve relatively high transmission capacities with relatively good spatial coverage. The high-band spectrum between 24 and 71 GHz served by millimeter waves in 5G has to be extended to 92 GHz, and the sub-THz band in the range from 92 GHz to 300 GHz should be considered for 6G, and later the THz spectrum up to 3 THz. In particular, the W-band from 92 - 114.25 GHz and the D-band from 130 - 174.8 GHz can be identified as 6G candidates. This would enable applications such as holograms and XR with ultra-high bit rates and ultra-low latencies. Concerning acceptable ranges, however, pencil-sharp beamforming, a massive number of antenna elements, and thus ultra-massive MIMO must be used. In addition, a 6G system must offer communications services for rural areas with long distances or massive obstacles, such as concrete ceilings between the base station and the terminals. This is only possible with the low-band spectrum below 1 GHz, possibly still with mid-band spectrum up to 7 GHz, although usable frequencies in this range are already being used for 2G, 3G, 4G, and/or 5G. The above considerations on frequency spectra for 6G are summarized illustratively in Figure 12.7 [266].

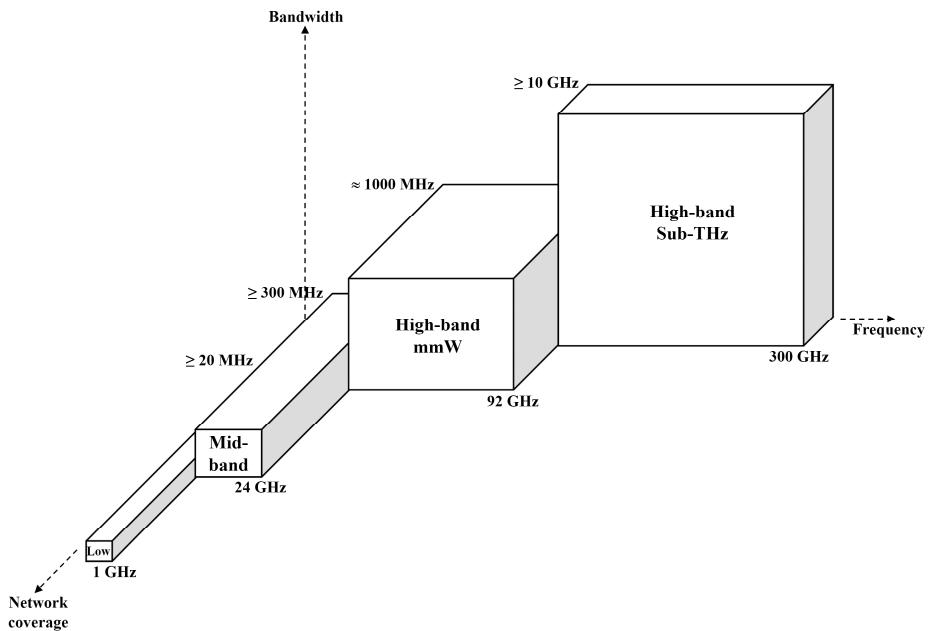


Fig. 12.7: Possible frequency spectrum for 6G [266]

One problem with the above-mentioned frequency ranges under consideration for 6G is that a considerable proportion of the frequencies is already used for other radio services, the low-band and lower mid-band spectrum up to 7 GHz primarily for the previous mobile communications generations, the upper mid-band spectrum up to 24 GHz for satellite-based and navigation services, among others. The high-band spectrum with mmW is also partially occupied by 5G; even the sub-THz range is already partly occupied by other radio services. This essentially limits 6G to the high-band spectrum for opening up completely new frequency ranges (spectrum exploration). In the low-band and lower mid-band spectrum, frequencies already used for previous mobile communications generations must be reallocated (spectrum refarming). However, this usually takes several years because of the networks in operation. In addition, efforts must be made to free up frequencies in the mid-band spectrum previously used for other purposes (spectrum clearing). These options are summarized in Table 12.4.

Tab. 12.4: Approaches to obtain licensed spectrum for 6G [266]

Frequency band	Exploring new frequency bands incl. spectrum sharing	Spectrum clearing	Spectrum refarming
Low-band spectrum			x
Lower mid-band spectrum	x	x	x
Upper mid-band spectrum	x	x	
High-band spectrum	x		

Making licensed frequency bands available for a new generation of mobile communications takes several years. Regarding coexistence with existing mobile communications systems, it is advantageous to start with previously unused frequency ranges when introducing a new generation and only reallocate frequencies in a second step.

In summary, the ITU-R conferences WRC-2023 and especially WRC-2027 are highly important for a 6G system.

12.4 6G Use Cases and Usage Scenarios

As in the case of 5G, the technical requirements for 6G are not based on the forecast technical possibilities but on the expected applications, i.e., use cases. This was also evident in the comments on Network 2030 (see Section 12.2). In addition, it has been formulated as a basic requirement that a future 6G system must be sustainable and, above all, should promote or massively strengthen sustainability in various industries and application areas.

As a result of this approach, numerous organizations, initiatives, and alliances have already addressed or are working on the question of use cases that require a network with performance characteristics that go well beyond 5G. In this context, the ITU-R Working WP 5D [260], the NGMN Alliance [204], the EU project Hexa-X [203], the North American Next G Alliance [202], the Chinese IMT 2030 (6G) Promotion Group [201], and the manufacturers Huawei [195] and Samsung [192] are particularly relevant. In this context, the cited documents frequently mention the following use cases, which are briefly explained below:

- Holographic-Type Communications (HTC)
- Extended Reality (XR)
- Multi Sensing
- Digital Twin
- Global seamless network coverage
- AI as a Service (AIaaS).

In other words, there is unanimous agreement worldwide on these use cases and that a 6G system is needed. In addition, the use cases

- interacting cobots and
- eHealth

are mentioned relatively frequently.

Holographic-Type Communications (HTC): Holography allows three-dimensional objects to be captured using the wave nature of light with coherence and interference and subsequently projected in free space. These can be people, things, or objects in general. If one captures sequences, one can also represent flows or changes. Three-dimensional viewing of the hologram on site is possible without 3D glasses. With a 6G system, such holographic data should not only be captured locally but also transmitted or streamed over the network. This is referred to as holographic-type communications. Using holographic display technology, enables novel applications such as holographic telepresence of people (digital twins) in a room for a meeting, the transmission of a holographic representation of a (hard-to-access) object, e.g., a machine to be repaired, telesurgery, or hands-on training across distances, e.g., the interaction of online training participants with ultra-realistic-looking objects [123; 191]. Remote rendering of high-resolution holograms over a mobile network provides a truly immersive experience. However, HTC leads to an enormous bandwidth requirement in the order of Tbit/s, even with image compression. In addition to frame rate, resolution, and color depth in two-dimensional video, volumetric data such as tilt, angle, and position are also critical to the quality of the hologram. HTC also requires extremely low latency for true immersion and high-precision synchronization of the corresponding data streams to reconstruct holograms [191].

Extended Reality (XR): XR refers to all technologies that deal with AR (augmented reality, the extension of the real world with virtual content, e.g., through AR glasses), VR (virtual reality, a completely virtual, computer-generated world, e.g., using VR glasses) and MR (mixed reality, a hybrid form between AR and VR). This allows virtual and real environments to be seamlessly merged. However, for an excellent immersive experience, compared to 2D video streaming, the videos are required with higher resolution, higher frame rate, greater color depth, and high dynamic range, resulting in a need for bit rates in the range of Gbit/s per end device, which is no longer feasible, especially at the edge of 5G NR radio cells. In addition, low latency and high reliability are essential for XR applications such as immersive gaming or industrial remote control [191]. Another challenge in supporting interactive XR experiences is the synchronous transport of multi-modal data streams (e.g., visual media, audio, and haptics) to and from different devices in a collaborative group operating the same XR application. Another important aspect is the support of real-time adjustments in the network in terms of user movement and actions to ensure that interactions with other users and objects appear realistic in terms of

placement and responsiveness. Enabling spatial interactions also requires fast accessibility to and easy integration of content with up-to-date and accurate representations of real/virtual environments from various content sources [260].

Multi Sensing: A human has five senses (sight, hearing, touch, smell, and taste) to perceive the external environment. However, current communications, including those at 5G, are limited to visual (text, image, and video) and acoustic (audio, voice, and music) media only. Including the senses of taste and smell would enable new user experiences and applications, e.g., in the food industry. In addition, haptic communication, i.e., communicating with touch, plays an increasingly important role and opens up a wide range of applications such as remote control and immersive games. The latter use cases have high requirements for low latency [191]. Sensing technology based on the measurement and analysis of radio signals, especially in the sub-THz range, opens up possibilities for high-precision positioning, ultra-high-resolution imaging, mapping and environment reconstruction, and gesture and motion recognition [260].

[190] gives an impression of the possibility of using THz signals for sensor technology. THz bands enable sensing and imaging, services for radio astronomy and Earth remote sensing, vehicle radars, chemical analysis, explosive detection, and moisture content analysis. In addition, THz signals can be used for wireless gas sensing, electronic smelling, and pollution monitoring, e.g., the concentration of certain greenhouse gases can be derived from a THz communication link. The same approach can be used to monitor chemical hazards in critical locations such as factories or laboratories. THz signals can also be used for medical imaging and material recognition, i.e., identifying a given object's shape and composition based on its spectral fingerprint. Therefore, THz technology supports e-health applications such as non-invasive tissue analysis and glucose concentration measurement. In addition, THz signals are strongly reflected by metal surfaces. This property can be used for security purposes, e.g., to detect the presence of weapons at airports. Finally, the directionality of THz links lends itself to supporting accurate localization and mapping services, supporting new applications such as high-resolution 3D mapping (projecting 3D sceneries), digital twin creation, immersive holographic telepresence, and interactive and cooperative robotics [190].

Digital Twin: A digital twin is a virtual copy of a real physical object or system, which in many cases requires high-precision sensing and high-bit-rate data transmission in real-time. The virtual copy is equipped with a wide range of features, information, and properties of the original object. As a result, testing and optimization can also be performed in the virtual environment. The results can be mirrored back to the real object or system for improvement. For example, if the object is a workpiece, manufacturing is also possible based on the optimized digital twin [191; 260].

Global seamless network coverage: A large part of the population in remote, sparsely populated, and rural areas still has no access to basic ICT services, result-

ing in a significant digital divide between different population groups on Earth. In addition, more than 70% of the Earth's surface is covered by water, so the growing number of maritime applications requires network coverage both on the surface and underwater. For both, global network coverage with sufficient capacity, acceptable quality of service (QoS), and affordable costs are still missing today. On the one hand, it is technically impossible for terrestrial networks to cover remote areas and extreme topographies such as oceans, deserts, and high mountains, and it is too costly to provide terrestrial communications services to sparsely populated areas. On the other hand, geostationary earth orbit (GEO) satellites are expensive, and their capacity is limited to a few Gbit/s per satellite, which makes their deployment feasible only for high-end users such as the maritime and aviation industries. These shortcomings of today's communications networks are to be mitigated in the first step with 5G by, among other things, integrating satellite-based base stations (see Section 9.3) with 3GPP Release 17. More comprehensively, this is to be solved with 6G by optimally exploiting the synergy of terrestrial networks, satellite constellations, and other airborne platforms to provide ubiquitous connectivity for global MBB users and large-scale IoT applications [191; 260].

AI as a Service (AlaaS): In the context of 6G, AI is seen on the one hand as a tool for automating the operation and/or optimizing the performance of the network itself, and on the other hand as a surrounding service provided by the mobile network for user applications. In short, this network service is described as AI as a Service (AlaaS) [204]. AlaaS can be used for classification and prediction in human-to-human and human-to-machine interactions based on criteria/features such as gestures, intonation, facial expressions, environmental sounds, touching of objects, etc. This service can be used by applications installed on user and IoT devices or in the network infrastructure to send requests for ML-based reasoning to the network, e.g., to other devices or to edge cloud hosts with already trained models. One conceivable example is supporting older people with movement or visual impairments who wear wearables with sensors to collect environmental data. This sensory data is then used to detect and identify objects, obstacles, and potential hazards to inform the user in advance and take proactive measures [203].

Interacting cobots: Increasingly close interaction between humans and robots is expected. Collaborative robots, so-called cobots, are used for this purpose. These networked cobots should be able to reliably read and interpret human actions and intentions, respond in a trustworthy manner, and thus support humans efficiently and safely. This would include use in industry but also as care assistants, exoskeletons, or adaptive wheelchairs. In this context, cobots could also form teams that solve tasks together, communicate to do so, and collaborate with humans on a group level [204].

eHealth: Based on communicating medical sensors (e.g., temperature, pulse, blood glucose level, blood pressure) and actuators (e.g., insulin pump, pacemaker, medication dispenser) worn on the body or implanted, i.e., with a highly available

and secure body area network (BAN), 6G could fundamentally transform healthcare to 24/7 monitoring of vital parameters via the network for both healthy and sick people. Accordingly, warnings can be issued, assistance can be provided, and therapy can be automated. Here, a virtual body double, the above-mentioned digital twin, could also be used advantageously. In addition, XR tools combined with haptic information such as surface, touch, actuation, motion, vibration, force, and audiovisual information can provide medical personnel with an immersive experience while using the insights of the digital twin [204].

In addition to the use cases and use case families already described, Table 12.5 provides an overview of the results of the EU project Hexa-X. On the one hand, it becomes clear that there are numerous other ideas for applications with 6G. This applies, for example, to smart cities, integrated micro-networks, and computing or security as a service. On the other hand, the important role of sustainability for 6G becomes obvious.

Tab. 12.5: Application categories and associated use cases [203]

Application category	Use Case
Sustainable development	eHealth for all Worldwide network coverage Earth monitoring Autonomous supply chains
Massive twinning	Digital twins for manufacturing Immersive Smart City Digital twins for sustainable food production
Immersive telepresence for enhanced interactions	Fully merged cyber-physical worlds Mixed reality co-design Immersive sports event Merged reality in games/work
From robots to cobots	Consumer robots AI partners Interacting and cooperative mobile robots Flexible manufacturing
Local trust zones for humans and machines	Precision healthcare Sensor infrastructure web 6G IoT micro-networks for smart cities Infrastructure-less network extensions and embedded networks Local network coverage for temporary use Small coverage, low power micro-networks in networks for production and manufacturing Automated public security
Conceivable further services	Compute-as-a-Service AI-as-a-Service (AlaaS)

Application category	Use Case
	AI-assisted Vehicle-to-Everything (V2X)
	Flexible device type change service
	Energy-optimized services
	Internet-of-Tags
	Security as a service for other networks

Sustainability in 6G means, on the one hand, that a 6G system is developed, manufactured, and operated as sustainably as possible, but on the other hand, that it is, above all, very well suited to support a wide range of environmental and social sectors in their sustainability efforts in the best possible way. To achieve this, it must, for example, offer solutions that contribute to achieving the SDGs (Sustainable Development Goals) of the United Nations (see Section 11.5) or help companies minimize their environmental impact. In particular, use cases that contribute to dematerialization (telepresence, e-health, XR), efficient use of resources (flexible manufacturing, AI partners), and optimization (energy-optimized services) can have a significant positive impact. The extreme power and global service coverage at 6G can help bridge the digital divide between people in different areas of the world, provide virtual yet realistic remote experiences, and provide means to monitor and address current and upcoming environmental issues [203]. The University of Oulu is working intensively on 6G and sustainability. The results to date are summarized in [229].

The addressed use cases, some of which are outlined above, can be summarized into six IMT-2030 deployment scenarios, as shown in Figure 12.8, according to [146]:

- Immersive Communication: This is an extension of eMBB from IMT-2020 with significantly higher data rates, increased mobility, lower latency if necessary, and higher reliability. Application examples are XR, holographic communication, and multi-sensory telepresence.
- Hyper Reliable and Low-Latency Communication: This represents an extension of URLLC from IMT-2020 for use cases with extremely stringent requirements for reliability, latency and often positioning, link density, and/or synchronicity, with expected severe consequences if not met. Typical use cases include industrial applications with full automation, remote motion control, interaction with and by robots, operation of drones, remote medical surgery, and transmission and distribution of electrical power.
- Massive Communication: This application scenario is an extension of mMTC from IMT-2020 and involves the ubiquitous connectivity of a huge number of devices and sensors, possibly with high mobility, security, and reliability requirements. Applications are able to capture, monitor, measure, and control the environment and gain knowledge about the real environment with the help of a huge number of objects. Application examples can be found in smart cities,

transportation, logistics, health, energy supply, environmental monitoring, agriculture, and generally in the IoT.

- Integrated AI and Communication: Compared to IMT-2020, this usage scenario includes entirely new applications that use distributed computing and AI. Typical use cases include network-assisted automated driving, the autonomous collaboration of devices for medical treatments, offloading of heavy computation operations across devices and the networks, creation and prediction using digital twins, network-assisted cobots, AI-assisted network automation, and AlaaS (AI as a Service) in general.

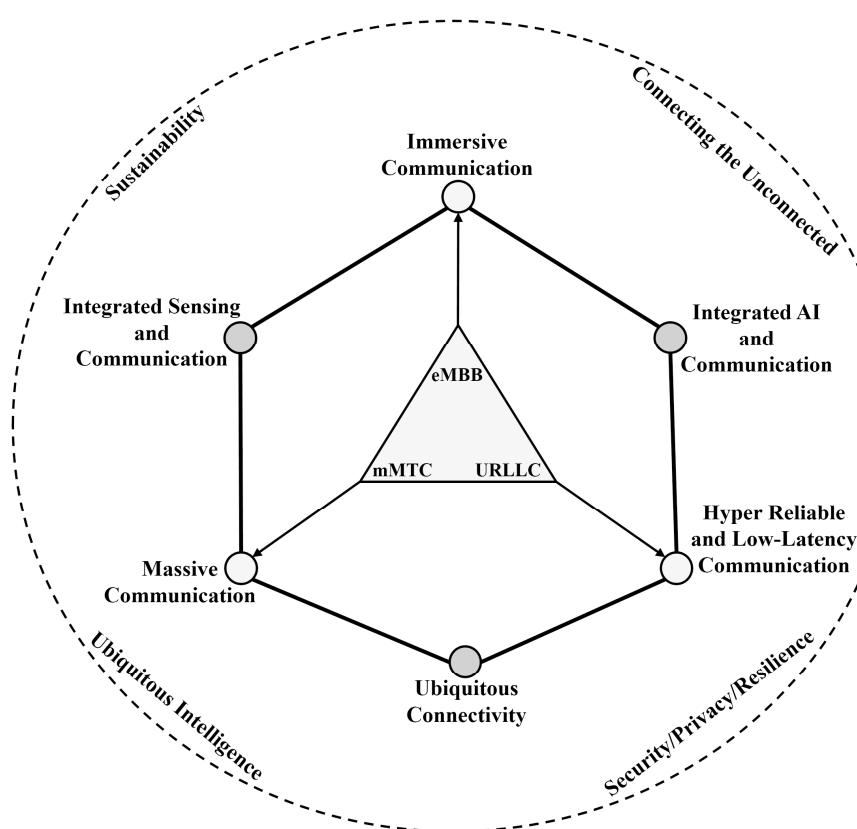


Abb. 12.8: IMT-2030 usage scenarios [146]

- Integrated Sensing and Communication: This application scenario supports use cases with multi-dimensional sensing that provide spatial information about objects, their movement, and their environment. In addition to communication capabilities, high-precision localization, object recognition, distance, speed,

and angle determination, among others, must be possible. Typical use cases include navigation, gesture and posture recognition, motion tracking, and providing spatial information for XR applications.

- Ubiquitous Connectivity: This usage scenario is intended to enhance connectivity with the aim of bridging the digital divide. One focus of this usage scenario is to address presently uncovered or scarcely covered areas, particularly rural, remote, and sparsely populated areas. Typical use cases include but are not limited to IoT and mobile broadband communication.

Figure 12.8 also takes into account four fields of application or functional areas that go beyond the five application scenarios or are to be taken into account by them [146]:

- Sustainability: This refers to the fact that an IMT-2030 system should comprehensively support applications to achieve the sustainability goals and, at the same time, be highly energy efficient, have minimal energy consumption, and have minimal impact on the environment during the manufacture, construction, operation, and maintenance of the network.
- Security/privacy/resilience: The protection of confidentiality, integrity, and availability of data, the user's self-determination over their data, and the network's resilience must be ensured in all deployment scenarios and corresponding use cases.
- Connecting the Unconnected: This considers connecting those previously un-reached by mobile network coverage, typically by integrating non-terrestrial network components, e.g., satellite-based base stations. This is also intended to maintain communications in the event of disasters.
- Ubiquitous intelligence: This field of application also relates primarily to rural areas that have been poorly reached to date, where the aim is to use IMT-2030 to drive forward digitization or overcome the digital divide, e.g., in payment transactions, logistics, and goods delivery.

12.5 6G Requirements

The concrete requirements for the technology of a 6G system can be derived from the 6G usage scenarios identified and the individual use cases summarized under them. The 6G IA proposes an interesting procedure in [189]. Among other things, it is based on the premise that a 6G system must address the SDGs (Sustainable Development Goals) of the United Nations for sustainable development (see Section 11.5). 6G networks will be an integral part of society and must, therefore, aim to cover societal challenges, problems, and needs and generate human-centric added value for society. For this reason, the first step in [189] is to start with the 17 UN SDGs and the 169 specific targets derived from them in accordance with [234]. They must be

interpreted for the ICT industry and 6G. A distinction must be made between the direct effects of 6G through, for example, material and energy consumption (sustainable 6G) and the indirect effects through the use of 6G (6G for sustainability). Taking this into account, so-called Social Key Values (SCVs) are derived in [189]. Starting from these, so-called Key Value Indicators (KVI) are derived based on use cases. These metrics are intended to make the impact of the KVs on society measurable or at least estimable. Depending on the concrete KVI, a different approach must be taken. E.g., measurements or estimates have to be made for use cases in a deployed network, experiments, interviews, questionnaire actions, etc., have to be conducted, or experts have to perform analyses. In the case of an exemplary global e-health service, positive influenced KVs would be societal sustainability, personal health, and protection from health harms. KVI would then be the number of e-health users, the average gain in healthcare access, the decrease in the number of hospital stay-overs, etc. For each KVI, so-called KV enablers are then determined, i.e., which key factors favor or impede the establishment of a KV. KV enablers in the e-health example would include global service coverage with MBB, low cost of connectivity, availability of secure cloud services, and the availability of appropriate end devices, subsidized if necessary. Subsequently, KVI are quantified with KPIs (Key Performance Indicator), i.e., concrete technical requirements for the future network. For the e-health service, KPIs would be, for example, the monthly cost of service for the user, the fraction of the world population covered by the e-health service, the energy consumption of the terminal devices, etc. According to [189], this very systematic approach to determining the requirements for a 6G system was carried out for the use cases "emergency response and warning systems" and "assistance from twinned cobots", among others. If this approach is applied consistently for all 6G use cases, a future 6G system could contribute much to sustainable development.

As explained in Section 12.4, numerous use cases for 6G have been and are still being developed. Based on these, the requirements for a 6G system are then developed, unfortunately, mostly without taking into account the sustainable approach outlined above. The requirements for IMT-2030 are discussed in the ITU-R Working Party 5D mentioned above and published in a draft [146] at the end of June 2023. Regionally active institutions such as the Beyond 5G Promotion Consortium (B5GPC) for Japan and the IMT-2030 (6G) Promotion Group for China, the globally active one6G Association, or manufacturing companies such as Samsung and Huawei provided contributions for this purpose. Accordingly, Table 12.6 summarizes the performance requirements (KPIs) for 6G and IMT-2030 published by the institutions and organizations mentioned. The trend for the target peak data rate under ideal conditions is toward 1 Tbit/s, while the bit rate achieved by the user at any time and any place with a probability of 95% is seen to be 1 Gbit/s and more. Regarding the targeted energy efficiency, the expected efficiency gains compared with 5G diverge widely, between two and one hundred times. In terms of spectrum efficiency, i.e.,

the value for bit/s/Hz, the values are 1.5 to three times those of 5G. There is a high level of agreement regarding the latency at the air interface of approximately 0.1 ms, the connectivity density of 10^7 devices/km 2 , and the system reliability that user data can be successfully transmitted with a probability of $1 - 10^{-7}$ or 99.99999%. The required traffic capacity values in Mbit/s/m 2 also differ significantly, between ten and thousand times, in relation to 5G. Again, there is widespread agreement on positioning accuracy at 1 cm and mobility support at up to 1000 km/h. In addition, almost all assume that the future 6G network with satellite communications is a global three-dimensional network supporting maritime applications on the world's oceans. In addition, Table 12.6 provides further, very ambitious requirements such as an end-to-end latency of 1 ms, a jitter at the air interface in the μ s range, an image resolution in the mm range, and a sensor battery lifetime of up to 20 years [257; 201; 152; 192; 195; 191].

Tab. 12.6: Performance requirements (KPIs) for 6G and IMT-2030

Requirements	B5GPC [257]	IMT-2030 (6G) [201]	one6G [152]	Sam- sung [192]	Huawei [195]	[191]
Peak data rate [Gbit/s]	100	100-1000	100	1000	1000	1000
User-experienced data rate [Gbit/s]	1	1-10		1	10-100	1
Energy efficiency (cf. 5G)	100x	20x	10x	2x	100x	10-100x
Spectrum efficiency (cf. 5G)		1,5-3x		2x		3x
Latency at air interface [ms]		0,1-1	0,1	0,1	0,1	<0,1
End-to-end latency [ms]	1			<1		
Connectivity density [devices/km 2]	10^7	10^7 - 10^8	10^7	10^7	10^7	10^7
Reliability	$1-10^{-7}$ or 99.99999%	$1-10^{-7}$	$1-10^{-7}$	$1-10^{-7}$	$1-10^{-7}$	$1-10^{-7}$
Traffic capacity (cf. 5G)		10-1000x	100x		1000x	100x
Positioning accuracy [cm]		1	1		50 outdoor, 1 indoor	
Imaging resolution [mm]	>1				1-3	
Jitter at air interface [μ s]		1		>1	+/-0,1	
Network coverage	3D incl. oceans	3D incl. oceans	3D incl. oceans		3D	
Sensor battery lifetime [years]					20	
Mobility [km/h]		1000		>>500		1000

As mentioned earlier, in the meantime, ITU-R WP 5D has specified capabilities and KPIs as target values of an IMT-2030 system in the form of a Draft New Recommendation (DNR) [146] and passed it on to Study Group 5 (SG 5) for final approval. Figure 12.9 shows the results. Thereby, Figure 12.9 distinguishes between the enhanced capabilities of IMT-2030 (6G) compared to IMT-2020 (5G) and completely new capabilities. The latter include, in particular, sensing-related capabilities, AI-related capabilities, and sustainability.

Interestingly, the ITU-R has not yet defined all the target values for the IMT-2030 capabilities. Some of them, like "peak/user experienced data rate", "spectrum efficiency", and "area traffic capacity", have still been designated as research target values. They are defined later in the preparation of the technical performance requirements paper.

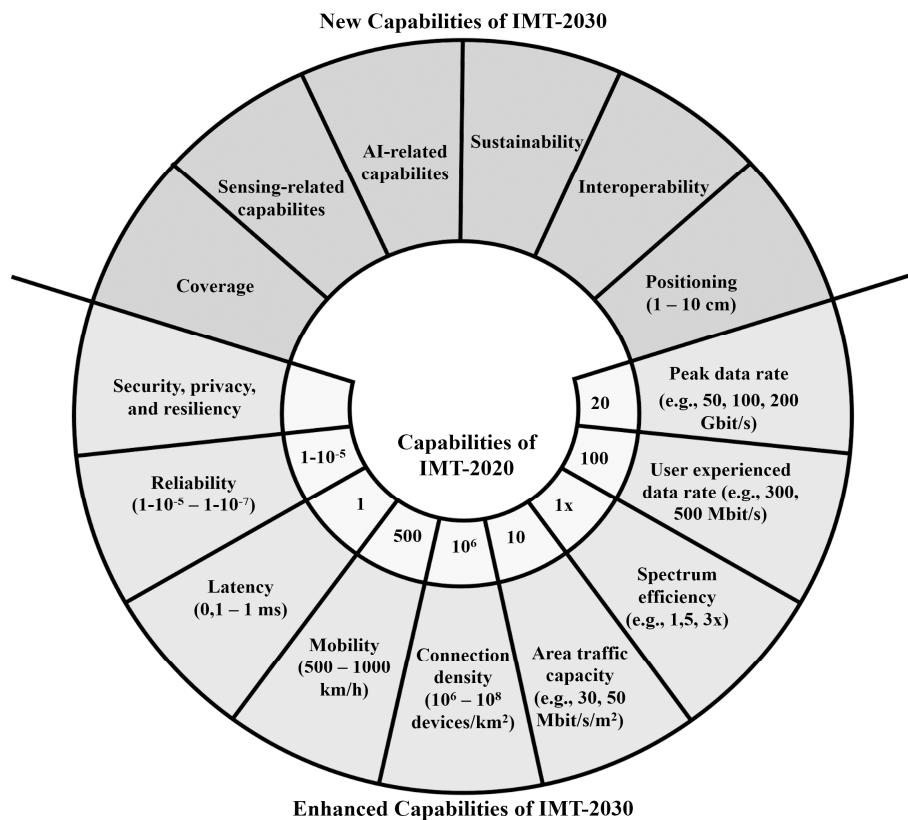


Fig. 12.9: Capabilities and key performance requirements (KPIs) for an IMT-2030 system (6G) compared with IMT-2020 (5G)

In addition, [192] defines architectural requirements:

- Communications and computing convergence: Due to the very high data rates, extremely low latency times, and distributed computing capabilities in the network, computationally intensive tasks can be shifted from insufficiently performing end devices to the network.
- Native AI: Appropriate AI functions must be embedded in the 6G system components to use AI to optimize the network, automate network operations, and provide AI functions to applications using the network (cf. AI-related capabilities in Figure 12.9).
- Integration of terrestrial and non-terrestrial network components: Concerning a truly global 3D network coverage, terrestrial fixed base stations must be provided on the one hand, and mobile base stations if required, on the other hand, base station functions in aircraft, HAPSs (High Altitude Platform Stations) and LEO (Low Earth Orbit), but also GEO (Geostationary Orbit) satellites (cf. Coverage in Figure 12.9).

[257; 192] adds the following additional requirements to the above architectural requirements:

- Very high level of synchronization in the network, e.g., for synchronization of cyber-physical systems (CPS) applications
- Autonomous network operation without human intervention (zero touch)
- Scalability of the network, that terminals can connect seamlessly with terrestrial and non-terrestrial base stations, also with other terminals, furthermore open interfaces (APIs) (cf. Interoperability in Figure 12.9)
- Network-, radio-based sensor technology (cf. Sensing-related capabilities in Figure 12.9)
- Highest requirements for trustworthiness, IT security, and privacy through a security-by-design approach to be able to trust the hardware and software used (cf. Security, privacy, and resiliency in Figure 12.9)
- System resilience, so that services and applications remain available or can be restored in the event of failures, even in the event of disasters (cf. Security, privacy, and resiliency in Figure 12.9)
- Sustainability: minimization of environmental impact through the use of environmentally friendly materials, good reusability, long-lasting devices, expandable software, modular hardware design, and in terms of CO₂ neutrality, use of renewable energies (cf. Sustainability in Figure 12.9).

12.6 Technologies for 6G and Network Architectures

Based on the use cases in Section 12.4 and the resulting requirements in Section 12.5, the question now is which technologies are required and which may already be

available for realization. The ITU-R, among others, attempts to provide initial answers to these questions in [260]. The trends and technical possibilities briefly discussed below are considered in this context.

Three scenarios for the use of AI-native communication in an IMT-2030 system are shown [260]: With an AI-native air interface, AI can be used advantageously for, e.g., symbol detection and decoding, radio channel property prediction, MAC layer design, and radio resource management. Among other things, an AI-native RAN operates in an automated manner, detects causes of faults and corrects them, optimizes energy consumption autonomously, provides on-demand computing capacity and functionalities, and optimizes the interaction of sensors and AI. In the third AI deployment scenario, the RAN itself supports AI services. For this purpose, computing and communication resources must be provided for distributed AI applications. As a result, autonomously operating systems with a large number of interacting IoT devices, for example, can be supported by AI algorithms.

Another promising technology for 6G is the integration of sensing and communication. This is called an integrated sensing and communication (ISAC) system. It enables radio-based object detection, range measuring, positioning, tracking, visualization, etc. In 5G, only positioning is supported. The additional functionalities mentioned for 6G benefit from the very high frequencies in the mmW and sub-THz range, the larger bandwidths, the higher base station density, the much larger antenna arrays, and the AI support.

The convergence of communications and computing architecture is also crucial for 6G. This enables split computing and distributed computing power, i.e., the joint use of computing resources on mobile terminals, base stations, MEC, and cloud servers. The restrictions resulting from the limited computing power on the terminals are overcome, and applications such as XR, mobile holograms, and digital twins become possible.

Sidelink communication, i.e., D2D communication of a UE across one or more UEs, plays an increasingly important role for cloud-based XR, tactile internet applications with remote motion control, UAVs, and SL-IIoT (Sidelink enhanced-Industry IoT), etc. This may require extremely high throughput in THz-, very low latency in sub-ms-, accurate positioning in cm-range, and low power consumption. D2D communication could also become essential to support the cooperation at a UE with peripheral devices such as computers, watches, smart glasses, autonomous vehicles, etc., through their direct sub-THz communication and cooperation to overcome the individually given resource constraints, e.g., the number of integrated antennas or UL bit rate [260]. The Next G Alliance goes one step further. Due to the ever-increasing density of mobile devices, local traffic and the possibilities for direct communication between UEs are increasing significantly. This leads to increasing sidelink transmission, even across multiple hops. This, in turn, results in greater network coverage. To be able to ensure reliability, scalability, and load balancing

despite this decentralized communication, mesh networks and even ad hoc networks with dynamic topology changes are being formed [183].

As mentioned in Section 12.3, a mix of different frequencies, including very high frequencies due to the large bandwidths required, must be used for the various applications. Optimal use of the available frequencies is possible with carrier aggregation (CA), distributed MIMO, central coordination of frequency use, and spectrum sharing, i.e., the joint use of available frequency spectra by 6G and 5G, for example. In addition, because of the very high data rates in the Tbit/s range, the hardware for signal processing, antennas, filters, amplifiers, mixers, oscillators, etc., must be implemented for radio signals in the sub-THz range.

In terms of IoT end devices, energy efficiency must be prioritized. Ideally, the end devices generate their operating energy (energy harvesting) from light, vibrations, temperature fluctuations, or radio waves. The latter is called backscattering, i.e., energy is extracted from received RF signals, and parts of the same signals are reflected with modulated data as a transmit signal. Another possibility for minimizing energy consumption, for example, for wearables, is operation as passive zero-energy devices with an on-demand triggered wake-up mechanism during data transmission. Backscattering can be beneficial for this as well. The network's energy consumption can be minimized by means of AI and clever network architecture, e.g., with a high density of base stations, as well as optimized signaling.

For real-time communication with extremely low latency, highly accurate time and frequency information must be provided in the network.

A future IMT-2030 system is heterogeneous and has to meet particular requirements regarding energy consumption, latency, and computing and storage resources. In addition, the extensive use of AI and the possible availability of quantum computers must be considered. Accordingly, it offers numerous novel attack possibilities. These need to be addressed by strong trustworthiness, enabling trust technology for security, privacy, and system resilience. This includes mechanisms for privacy protection, especially in the RAN with distributed ledger techniques and federated learning, location-based cryptography in the RAN with the geographic location as proof of identity for post-quantum security, and adaptive security procedures using coding and anomaly detection on the physical layer.

Significant technological advances in radio interfaces are also crucial for an IMT-2030 system. First of all, the modulation methods must be mentioned here. For bit rates of 1 Tbit/s and more, improvements in QAM will no longer be sufficient. For example, new modulation methods with pulse shaping are being discussed. Another important topic for future radio interfaces is channel coding for extreme performance, incl. low power consumption, and for multiple use cases. Necessary progress here concerns both the integrated circuits and the algorithms. Previously based on OFDM, the waveform design must also achieve significant improvements. Last but not least, the multiple access method used must be optimized. In addition to the OMA (Orthogonal Multiple Access) method used to date in the form of

OFDMA, NOMA (Non Orthogonal Multiple Access) could also play a role in 6G with the possibility of simultaneous use of the same frequency resources by several terminals, including with Massive MIMO with highly directional beams.

In combination with the mentioned transmission technologies for the radio interfaces, advanced antenna technologies are, of course, also required. There is potential here, especially in MIMO systems, since their information-theoretical limits are far from being reached in practice. E-MIMO (Extreme-MIMO) should be mentioned here as a further development of Massive MIMO with larger antenna arrays of hundreds to thousands of phase-synchronized antennas and new materials. This could also be used to determine precise positions and visualize objects in the environment, including creating digital twins. The use of passive reflective elements with electronically controllable RF operation is referred to as holographic MIMO. E-MIMO could be further optimized by Distributed E-MIMO, i.e., the distribution of E-MIMO antennas over a larger area to improve coverage and reduce energy consumption due to proximity to users. From their point of view, the network then appears to have no radio cell boundaries concerning DL data transmission. Therefore, one speaks of cell-less or cell-free communication. A further improvement of Distributed E-MIMO is possible using AI-based techniques for the numerous optimization parameters.

Another approach due to scarce frequency resources is in-band full-duplex transmission with Self Interference Cancellation (SIC) technology.

Further possibilities for optimal use of the very scarce frequency resources, especially below 6 GHz, exist through the best possible utilization of the spatial dimensions for radio transmission. First, reconfigurable intelligent surfaces (RIS) should be mentioned here. Despite the use of mainly passive elements, integrated, e.g., in walls, they allow dynamic control over the radio environment by adjusting channel parameters such as phase, amplitude, frequency, and polarization through the tunable scattering of electromagnetic waves. Also worth mentioning in this context is Holographic Radio (HR) with exploitation of spectral aperture and interference for spatial and spectral multiplexing, and Orbital Angular Momentum (OAM) with a twisting of the propagating beams so that they become orthogonal to each other.

Regarding frequency ranges, applications requiring extremely high data rates, low latency, high-resolution sensing and imaging, and high-precision positioning are use cases for sub-THz and THz communications. Disadvantages include high propagation losses due to the very high frequencies, attenuating interactions of the electromagnetic waves with the atmosphere, rainfall, and leaves, and sensitivity to obstacles such as walls, vehicles, and people. Solutions for this can be extremely narrow beams, so-called pencil-beams, with the advantage that the antenna arrays require very little space at such high frequencies. This must be accompanied by the challenging development of correspondingly powerful and, at the same time, very energy-efficient THz transceivers.

In future IMT-2030 systems, a UE positioning accuracy in the cm range with a latency of a few tens of ms should be given. THz technology should significantly contribute to this, if necessary, in combination with AI.

As an alternative or supplement to sub-THz communication, mainly for indoor use, so-called Visible Light Communication (VLC) is considered. Since the interference is very low, a high cell density is possible. Since no walls can be penetrated, security and privacy can be easily ensured. Hundreds of THz of unlicensed bandwidth are available, and bit rates in the range of 10 Gbit/s are possible. The disadvantage is that line-of-sight (LOS) must always be guaranteed for transmission [246; 107].

The RAN is assumed to be a user-centric cell-free network. Each RAN resource – frequency bands, sub-channels, processing time – can be partitioned according to the QoS requirements of the various packet flows or applications and thus divided into slices. Adaptive and, with AI, intelligent RAN slicing can thus be supported. Each user may have their own virtual network or slice. In addition, it is assumed that the RAN is also service-based on the basis of micro-services (see Section 8.3). The RAN architecture should be designed to support, among other things, the convergence of data, operations, information and communication technologies, especially also because of AI usage and AI provisioning. In addition, a 6G RAN architecture should be as simple and unified as possible and separate signaling and user data.

Another approach to the RAN is the digital twin network (DT-RAN), which can be used to study, simulate, and optimize a RAN at runtime without disruptive effects on the physical network.

An IMT-2030 system must also provide mechanisms and technologies for the seamless interconnection of terrestrial and non-terrestrial networks with satellites, HAPSs (High Altitude Platform Stations), and UASs (Unmanned Aircraft Systems). NFV, SDN, network slicing, edge computing, and free-space optical communications are key technologies for this. However, the highly dynamic network topology, the different operating environments, and the very different delay times must be considered.

An ultra-dense RAN must be provided because of the very high data rates, UE density, power and spectrum efficiency, and coverage. For this purpose, connecting the numerous RUs by radio may be cheaper than fiber.

Regarding effort and costs, future RAN infrastructures should be shared by several operators (RAN sharing). For this purpose, issues such as user privacy, QoS guarantees, slice usage, and secure data storage, e.g., using blockchain technology, need to be addressed [260].

Although the work on 6G is still mainly at the research stage, and we are talking about regular network operation from around 2030, surprisingly, almost all publications on this subject assume the continued use of IP technology. Only in the work on Network 2030 (see Section 12.2) and in the Finnish 6G Flagship project [205]

the need for an evolution of IP is seen. One step proposed in [205] is to add a contract field between the header and payload to the IP packet. This would allow quasi-contractual packet delivery requirements to be specified and transmitted along with it, such as a latency guarantee for timely packet delivery. A second proposed extension relates to addressing to support mobility better paired with location-based real-time information or to be able to embed geocoordinates in the address format for satellite communication. Third, the data in the payload field should not all be handled the same way but according to their importance to the application, which is relevant, for example, for the different video stream frames (see Section 12.2). For such an IP evolution to a new IP packet format, backward compatibility with the existing IP must be ensured [205].

Based on the use cases (see Section 12.4), the requirements derived from them (see Section 12.5), and the technologies explained above that are already available or still to be researched or developed, a possible network architecture for a 6G or IMT-2030 system can be worked out. In the EU project Hexa-X, eight architecture principles were derived in the first step, which are mentioned and briefly explained below [246]:

- Exposure of capabilities: Existing and new network functions should be made visible and accessible externally so that applications can use them for an end-to-end view. Based on this, an application can then, for example, initiate a defined latency response, use performance predictions, or initiate a predictive orchestration (cf. NEF for 5G, see Section 8.2).
- Designed for (closed-loop) automation: The management of the network and the services should be fully automated and, as far as possible, without human interaction. Observation and analysis of network behavior should be directly fed back to optimizing the affected network functions. This is realized with the help of AI.
- Flexibility to different topologies: The network must be able to adapt to new circumstances, such as the integration of new subnets, non-public networks, autonomous networks, mesh networks, or RAN technology with new frequency bands without impacting performance and current end-to-end services.
- Scalability: The system architecture must enable a wide range of network sizes, from very small to very large networks. To achieve this, it must be possible to scale the network resources, such as the cloud platforms, accordingly.
- Resilience and availability: The separation of CP and UP functions coupled with multi-connectivity should ensure that the services provided and the underlying network infrastructure are highly available and fail-safe. This also means that if a subnetwork loses its connectivity to the overall network, it can connect to another subnetwork, thus avoiding a single point of failure.
- Exposed interfaces are service based: Interfaces for accessing network functions from outside must be cloud-native to be embedded directly in cloud environ-

- ments. In particular, it must be possible to reuse network functions as micro-services and add new NFs in plug-and-play.
- Separation of concerns of network functions: Different network functions should have no or only minimal dependencies and communicate with each other only via their APIs so that they can be developed, modified, and exchanged independently of each other.
 - Network simplification in comparison to previous generations: The goal must be for the network architecture to minimize complexity in the network, among other things, by using cloud-native RAN and CN functions, a few external interfaces, and a few parameters to be configured.

As already indicated in the architecture principles "flexibility to different topologies", "scalability", and "resilience and availability", a future 6G network is not a monolithic network with RANs and CNs but an overall system made up of many individual interconnected networks of different sizes and functionality. This is referred to as a network of networks [246]. Such a new network architecture supports unlimited connectivity with global coverage, enabling and providing big data and AI-supported applications. This architectural approach combines terrestrial and non-terrestrial subnets, D2D communications, mesh networks, campus/private networks, and locally formed ad hoc networks to provide flexible multi-connectivity. Flexible topologies and subnetwork integrations are the key to globally available services, low latency, high reliability, and security, even for services with extreme requirements [246]. With the network of networks, the modularity principle for fulfilling a wide variety of network requirements is elevated to the network level. In this context, the Finnish 6G Flagship project talks about ManyNets as the basis of a 6G system [205]. Depending on the end devices and applications to be supported, these subnets with different performance features can be operated autonomously stand-alone or in conjunction with a 6G wide-area network, as required. Network functions and services can be identified in the local subnet or the WAN, used and moved between both [245].

Figure 12.10 shows the 6G network architecture characterized above. Here, terrestrial subnets and NTN with satellites, HAPSS (High-Altitude Platform Station, e.g., airship/Zeppelin or aircraft), and/or drones, as well as campus, multi-hop and mesh networks or even the communications part of a high-speed rail network are interconnected. The network is spanned in three dimensions. In the case of terrestrial subnets, macrocells operate at lower frequencies, or microcells or femtocells operate at mmW or in the sub-THz range, depending on requirements, including spatial coverage and bit rates. The latter seems to be more suitable for in-house use with a defined environment due to the physical limitations. In principle, as indicated in Figure 12.10 and mentioned above, a cloud- and service-based architecture is assumed for both the actual network functions and the network management and

orchestration. In particular, the SBA approach from 5G (see Section 8.3) should be extended to the 6G RAN [246].

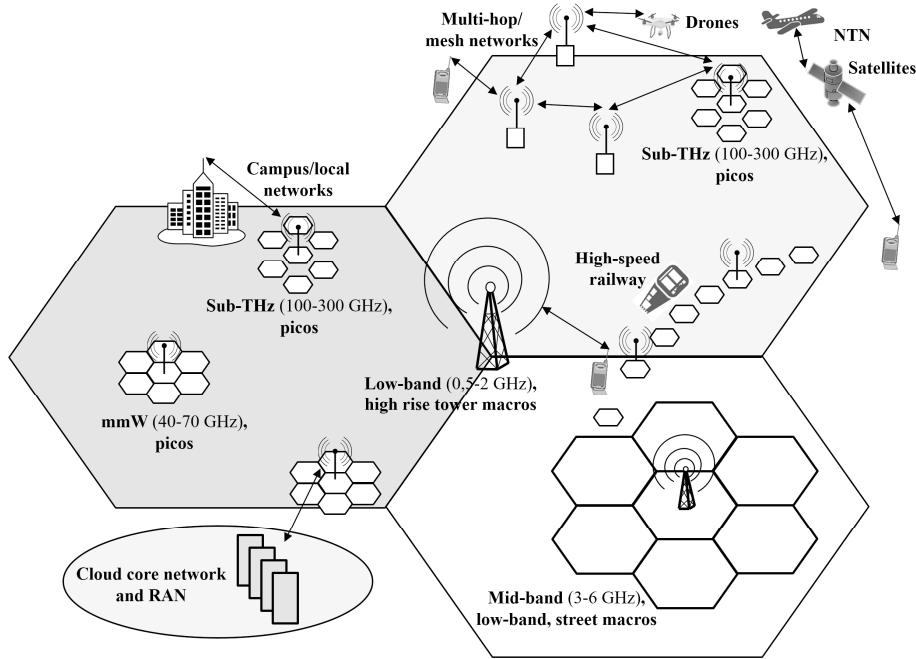


Fig. 12.10: 6G Architecture with Networks of Networks [246]

About the orchestration of network resources and functions, [246] assumes a concept of continuous orchestration, as shown in Figure 12.11. This means that orchestration is performed not only in the core network but also in the (R)AN or more generally in the edge network and private networks. This means that orchestration takes place not only in the core network, as in the case of 5G, but also in the transport network, in the (R)AN or more generally in the edge network, as well as in private networks, IoT subnets, and even the end devices of the users, such as smartphones and even wearables. The latter devices, which may be very heterogeneous, are summarized in Figure 12.11 under the term extreme edge. This is called a continuum of device edge cloud management [246]. The computing power available in the core, edge and extreme edge areas can also be used for the above-mentioned split computing using high-bit-rate communication within the system and with the UEs.

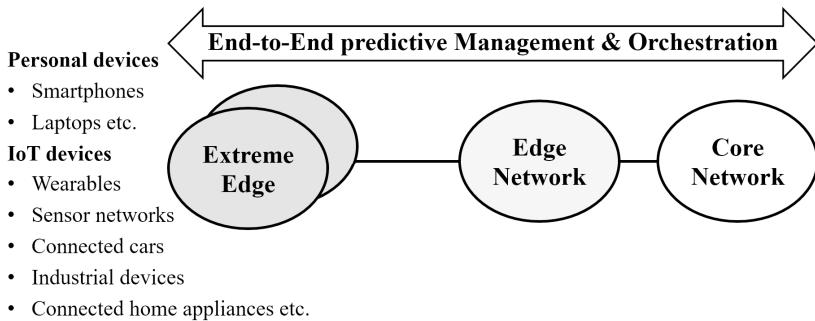


Fig. 12.11: Continuum Orchestration in 6G [246]

Most of the 6G subnets mentioned are in the edge area. These are often subnetworks tailored to specific applications or industries (verticals) with a rather small size. [245] talks about so-called special purpose networks (SPN) in these cases. They can operate independently but also benefit from the connection to the WAN. Examples of such 6G SPNs, each with its own requirements, are Industry 4.0 networks, a network of cooperating robots or communicating drones, a vehicle network for engine control, antilock braking, and driver assistance, among others, and a body area network (BAN) with sensors, pacemakers, insulin pumps, etc. [245].

Based on Figures 12.10 and 12.11 and the associated descriptions, an attempt was made to sketch a possible overall view of a 6G system in Figure 12.12, following [267]. It is assumed that attention is paid to maximum energy efficiency and that trust, security and privacy by design are taken into account from the very beginning. Virtualization plays a significant role in all network partitions. The access network uses wireline connections, a wide variety of terrestrial up to the sub-THz range, and non-terrestrial radio access. This creates a three-dimensional network with global network coverage. The radio interfaces, which operate in very high-frequency ranges, offer comprehensive, integrated sensor functions. As a result of the high-bit-rate and thus real-time-capable networking, the high-performance computing resources distributed in the network can be used for the network functions themselves but also for applications with distributed processing in the sense of Compute as a Service (CaaS) or AI as a Service (AIaaS). A 6G system would thus represent a distributed AI and a computing platform. According to Figure 12.12, the SBA approach is consistently extended to the RAN for 6G, with a separation into user plane (UP) and control plane (CP) network functions and disaggregation wherever possible and appropriate. Subnets, especially special purpose networks for special use cases, will play a much more important role than today. They can be connected directly to the 6G core network or via the access network or operate autonomously. The applications use the resources of the network up to the end devices. Accordingly, orchestration and management also take place across the entire network. Figure 12.12 struc-

tures the mentioned network functions according to [267] in three layers: infrastructure and cloud layer, network functions layer and application layer.

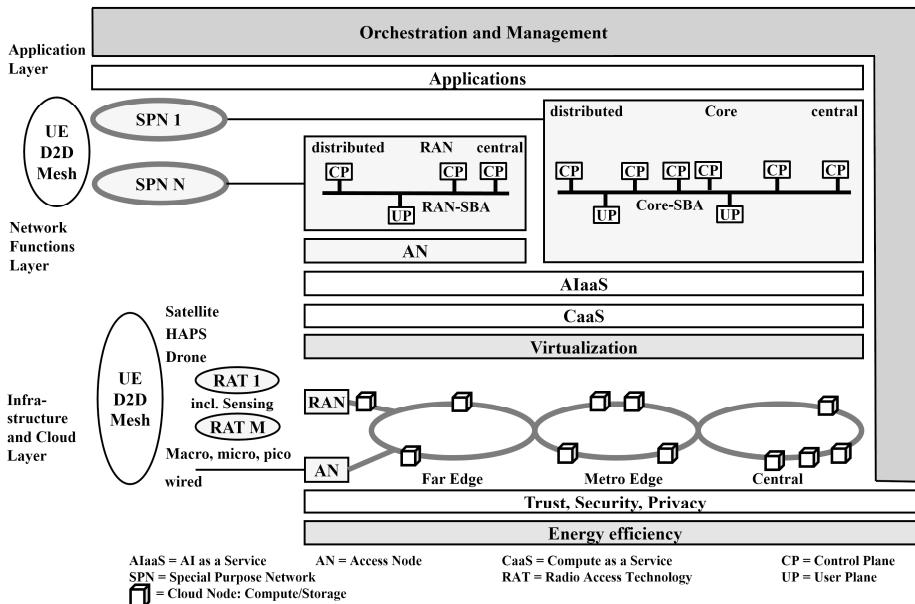


Fig. 12.12: 6G system in an overall view

In conclusion to this 6G overall system overview, the innovations and advantages of 6G compared to 5G are summarized below. According to the current status, the capabilities in the usage scenarios are significantly expanded; eMBB becomes Immersive Communication, URLLC becomes Extreme Communication, and mMTC becomes Massive Communication. As a result of the added high frequencies up to sub-THz ranges, there is a new application scenario, Integrated Sensing and Communication with integrated, comprehensive sensor technology. Another very innovative application category is Integrated AI and Communication, which enables not only AI-supported operation and optimization of the 6G network itself but also applications for users, i.e., AlaaS that provides an AI platform that is available everywhere and also has access to network-specific information such as locations. Compute as a service or split computing could also be offered, i.e., using distributed computing power in the network, e.g., by insufficiently performant UEs. For the first time, 6G should make it possible to cover inaccessible regions and even the world's oceans with high-performance communications technology by integrating satellite, HAPS, and/or drone-supported base stations. For the RAN, it is planned, on the one hand, to significantly extend the mmW frequency range upwards and, on the other hand,

to make sub-THz frequencies usable for the first time. This requires, among other things, the use of Extreme MIMO and possibly RIS systems. Because of its advantages, the SBA approach is also being applied in the RAN. Beyond the campus networks introduced with 5G, 6G will, from today's perspective, have numerous subnets operating stand-alone and in interconnection with the WAN and each other. They can be specialized and provide functions or features beyond the 6G WAN capabilities, e.g., in the case of body area networks. These are then so-called special purpose networks. If this approach is applied, a network of networks with orchestration down to the SPNs emerges. In addition to very powerful end devices and the RedCap and IoT devices, there will be a new class of very simple end devices, especially sensors, which generate their operating energy via energy harvesting. Above all, the research and development of 6G is focused on sustainability, both in terms of the 6G system itself but also in terms of its contribution to many application areas and industries. May this be not only the intention but come true in the real-world deployment of future 6G networks.

Abbreviations

A

a	Attributes
AAA	Authentication, Authorization and Accounting
ABF	Application-aware data Burst Forwarding
A-CPI	Applications-Controller Plane Interface
AF	Application Function
AGF	Access Gateway Function
AGV	Automated Guided Vehicle
AI	Artificial Intelligence
AlaaS	AI as a Service
AKA	Authentication and Key Agreement
ALG	Application Layer Gateway
AMBR	Aggregate Maximum Bitrate
AMF	Access and Mobility Management Function
AMR	Adaptive Multi-Rate
AN	Access Network
AN	Access Node
API	Application Programming Interface
App	Application
APT	Advance Persistent Threat
AR	Augmented Reality
ARP	Allocation and Retention Priority
ARPF	Authentication credential Repository and Processing Function
ARQ	Automatic Repeat Request
AS	Access Stratum
AS	Application Server
ASN	Abstract Syntax Notation
ATCF	Access Transfer Control Function
ATGW	Access Transfer Gateway
ATM	Asynchronous Transfer Mode
ATSSS	Access Traffic Steering, Switching and Splitting
AuC	Authentication Center
AUSF	Authentication Server Function
AVP	Attribute-Value Pair

B

B2BUA	Back-to-Back User Agent
B5GPC	Beyond 5G Promotion Consortium
BAKOM	Bundesamt für Kommunikation

BAN	Body Area Network
BAR	Buffering Action Rule
BBE	Beyond Best Effort
BBE & HPC	Beyond Best Effort & High Precision Communications
BBF	Broadband Forum
BBU	Base Band Unit
BfS	Bundesamt für Strahlenschutz
BGCF	Breakout Gateway Control Function
BGP	Border Gateway Protocol
BGP-FS	Border Gateway Protocol-Flow Spec
BS	Base Station
BICC	Bearer Independent Call Control
BMBF	Bundesministerium für Bildung und Forschung
BNG	Broadband Network Gateway
BNetzA	Bundesnetzagentur
BSC	Base Station Controller
BSI	Bundesamt für Sicherheit in der Informationstechnik
BSS	Business Support System
BTS	Base Transceiver Station

C

c	Connection Data
CA	Carrier Aggregation
CAP	CAMEL Application Part
CaaS	Compute as a Service
CBRS	Citizens Broadband Radio Service
CDN	Content Delivery Network
CER	Capabilities-Exchange-Request
CHF	CHarging Function
CN	Core Network
CriC	Critical Communications
CP	Control Plane
CP	Cyclic Prefix
CPE	Customer Premises Equipment
CPN	Customer Premises Network
CP-OFDM	Cyclic Prefix-OFDM
CPS	Cyber Physical System
C-RAN	Cloud-RAN
C-RAN	Centralized-RAN
CRC	Cyclic Redundancy Check
CRM	Customer Relationship Management

CS	Call Server
CS	Circuit-Switched
CSAI	Connectivity and sharing of pervasively distributed AI data, models and knowledge
CSCF	Call Session Control Function
CSFB	Circuit Switched Fallback
CU	Control Unit
CW	Codeword

D

D2D	Device-to-Device
D/A	Digital/Analog
DC	Dual Connectivity
D-CPI	Data-Controller Plane Interface
DDoS	Distributed Denial of Service
DHCP	Dynamic Host Configuration Protocol
DL	Down Link
DM-RS	Demodulation References Signal
DN	Data Network
DNN	Data Network Name
DSC	Diameter Signaling Controller
DSCP	Differentiated Services Code Point
DSL	Digital Subscriber Line
DSS	Dynamic Spectrum Sharing
DSS1	Digital Subscriber Signalling system no. 1
DS-TT	Device-Side TSN Translator
DT	Digital Twin
DU	Distributed Unit

E

e	evolved
E2E	End-to-End
EAP	Extensible Authentication Protocol
EBS	Educational Broadband Service
E-CSCF	Emergency-CSCF
EDGE	Enhanced Data Rates for GSM Evolution
EDR	Emergency and disaster rescue
EIR	Equipment Identification Register
EHD	Extrem HD-Video
ELPC	Extremely Low-Power Communications
eMBB	Enhanced Mobile Broadband

EMF	Electromagnetic Fields
E-MIMO	Extreme-MIMO
EMS	Element Management System
eNB	evolved NodeB
EN-DC	E-UTRA-NR Dual Connectivity
EPC	Evolved Packet Core
ePDG	Evolved Packet Data Gateway
EPS	Evolved Packet System
ERLLC	Extremely Reliable and Low-Latency Communications
ERP	Effective Radiated Power
ETSI	European Telecommunications Standards Institute
E-UTRAN	Evolved-UTRAN
eV2X	Enhancement of Vehicle-to-Everything
EVS	Enhanced Voice Services

F

FAR	Forwarding Action Rule
FCAPS	Fault Management, Configuration, Accounting, Performance and Security
FDD	Frequency Division Duplexing
FEC	Forward Error Correction
FeMBB	Further-enhanced Mobile Broadband
FFT	Fast Fourier Transformation
FG	Focus Group
FG	Forwarding Graph
FMC	Fixed Mobile Convergence
FMIF	Fixed-Mobile Interworking Function
FN	Fixed Network
FN	Future Network
FN-RG	Fixed Network-Residential Gateway
F-OFDM	Filtered-OFDM
FR	Frequency Range
FRMCS	Future Railway Mobile Communication System
FSS	Fixed Satellite Services
FWA	Fixed Wireless Access

G

GEO	Geo-stationary Earth Orbit
GERAN	GSM/EDGE Radio Access Network
GGSN	Gateway GPRS Support Node
GMSC	Gateway-MSC
GP	Guard Period

GRPS	General Packet Radio Service
GPS	Global Positioning System
gPTP	generalized Precision Time Protocol
GR	GRPS Register
GRE	Generic Routing Encapsulation
GSM	Global System for Mobile Communications
GSMA	Groupe Speciale Mobile Association
GST	Generic Slice Template
GTP-U	GPRS Tunneling Protocol-User plane

H

HAPS	High Altitude Platform Station
HARQ	Hybrid Automatic Repeat Request
HBF	Holographic Beamforming
HE	Home Environment
HEO	High Elliptical Orbit
Hetnet	Heterogeneous Network
HF	High Frequency
HLR	Home Location Register
HLS	High Layer Split
HPLMN	Home PLMN
HR	Holographic Radio
HSD	Huge Scientific Data pplications
HSDPA	High Speed Downlink Packet Access
hSEPP	home SEPP
HSM	Hardware Security Module
H-SMF	Home-Session Management Function
HSPA	High Speed Packet Access
HSS	Home Subscriber Server
HSUPA	High Speed Uplink Packet Access
HTC	Holographic-Type Communications
HTTP	Hypertext Transfer Protocol
HTTP/2	Hypertext Transfer Protocol Version 2
HW	Hardware

I

I2RS	Interface to the Routing System
IA	Industry Association
IaaS	Infrastructure as a Service
IAB	Integrated Access und Backhaul
IANA	Internet Assigned Numbers Authority

IARC	International Agency for Research on Cancer
IBCF	Interconnection Border Control Function
ICIC	Inter-cell Interference Coordination
ICMP	Internet Control Message Protocol
ICNIRP	International Commission on Non-Ionizing Radiation Protection
I-CSCF	Interrogating-CSCF
ICT	Information and Communication Technology
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IFFT	Inverse Fast Fourier Transformation
IIoT	Industrial IoT
IMS	IP Multimedia Subsystem
IMT	International Mobile Telecommunications
INAP	Intelligent Network Application Part
IoE	Internet of Everything
ION	Intelligent operation network
IP	Internet Protocol
IPsec	IP Security
ISAC	Integrated Sensing and Communication
ISC	IMS Service Control
ISDN	Integrated Services Digital Network
ISG	Industry Specification Group
ISUP	ISDN User Part
ITU	International Telecommunication Union
ITU-R	ITU-Radiocommunication Sector
ITU-T	ITU-Telecommunication Standardization Sector
I-UPF	Intermediate-UPF

J

J	Joule
JSON	JavaScript Object Notation

K

KPI	Key Performance Indicator
KV	Key Value
KVI	Key Value Indicator

L

LDHMC	Long-Distance and High-Mobility Communications
LDPC	Low Density Parity Check
LED	Light-emitting Diode

LEO	Low Earth Orbit
LI	Lawful Interception
LIA	Location-Info-Answer
LINP	Logically Isolated Network Partition
LIR	Location-Info-Request
LIS	Large Intelligent Surfaces
LL	Low Layer
LLS	Low Layer Split
LOS	Line of Sight
LPWAN	Low Power Wide Area Network
LRF	Location Retrieval Function
LTE	Long Term Evolution
LTE-M	LTE for Machines

M

m	Media Descriptions
M2M	Machine to Machine communications
MAA	Multimedia-Auth-Answer
MAC	Medium Access Control
MANO	Management and Orchestration
MAP	Mobile Application Part
MAR	Multi-Access Rule
MAR	Multimedia-Auth-Request
MBB	Mobile Broadband
MBMS	Multimedia Broadcast and Multicast Services
MC	Mission Critical
MCPTT	Mission Critical Push To Talk
MCU	Multipoint Control Unit
MDT	Minimization of Drive Tests
ME	Mobile Equipment
MEC	Multi-access Edge Computing
MEO	Medium Earth Orbit
METIS	Mobile and wireless communications Enablers for the Twenty-twenty Information Society
MGC	Media Gateway Controller
MGCF	Media Gateway Control Function
MGW	Media Gateway
MIMO	Multiple Input Multiple Output
MIoT	Massive Internet of Things
ML	Machine Learning
MM	Mobility Management

MME	Mobility Management Entity
mMTC	Massive Machine Type Communications
MMTel	Multimedia Telephony service
mmW	millimeter Wave
MPEG	Moving Picture Experts Group
MPLS	Multiprotocol Label Switching
MPS	Multimedia Priority Service
MPTCP	Multipath TCP
MR	Multi-Radio
MR	Mixed Reality
MRB	Media Resource Broker
MRF	Media Resource Function
MRF	Multimedia Resource Function
MRFC	Multimedia Resource Function Controller
MRFP	Multimedia Resource Function Processor
MS	Mobile Station
MS	Media Streaming
MSC	Message Sequence Chart
MSC	Mobile Switching Center
MTC	Machine-Type Communications
MTP	Message Transfer Part
MTSDT	Mobile Terminated-Small Data Transmission
MVNO	Mobile Virtual Network Operator

N

N3IWF	Non-3GPP InterWorking Function
N5CW	Non-5G-Capable over WLAN
NaaS	Network as a Service
NAPT	Network Address and Port Translation
NAS	Non Access Stratum
NB	Narrowband
NBI	Northbound Interface
NB-IoT	Narrowband-IoT
NCC	Network and computing convergence
NE-DC	NR-E-UTRA Dual Connectivity
NEF	Network Exposure Function
NEST	Network Slice Type
NETCONF	Network Configuration Protocol
NF	Network Function
NFV	Network Functions Virtualisation
NFVI	NFV Infrastructure

NFVIaaS	NFVI as a Service
NFVI-POP	NFV Infrastructure-Point of Presence
NFVO	NFV Orchestrator
NG	Next Generation
NGAP	NG Application Protocol
NGC	Next Generation Core
NGEN-DC	NG-RAN E-UTRA-NR Dual Connectivity
NGMN	Next Generation Mobile Networks
NGN	Next Neneration Networks
NIC	Network Interface Card
NIR	Non-Ionizing Radiation
NISV	Verordnung über den Schutz vor nichtionisierender Strahlung
NLOS	Non Line of Sight
NOMA	Non-orthogonal Multiple Access
NPN	Non-Public Network
NR	New Radio
NR-DC	NR-NR Dual Connectivity
NRF	Network Repository Function
NSA	Non-Standalone
NSH	Network Service Header
NSI	Network Slice Instance
NSSAAF	Network Slice Specific Authentication and Authorization Function
NSSAI	Network Slice Selection Assistance Information
NSSF	Network Slice Selection Function
NSSI	Network Slice Subnet Instance
NTN	Non Terrestrial Network
NTN-GW	Non-Terrestrial Network-Gateway
NWDAF	Network Data Analytics Function
NW-TT	Network-Side TSN Translator

O

OAM	Operation, Administration and Maintenance
OAM	Orbital Angular Momentum
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
ONF	Open Networking Foundation
ONIR	Ordinance on Protection from Non-ionising Radiation
OS	Operating System
OSS	Operations Support System
OTT	Over The Top
OVSDB	Open vSwitch Database Management Protocol

P

PAL	Priority Access License
PBCH	Physical Broadcast Channel
PCEP	Path Computation Element Communication Protocol
PCF	Policy Control Function
PCRF	Policy and Charging Rules Function
P-CSCF	Proxy-CSCF
PCU	Packet Control Unit
PDCCH	Physical Downlink Control Channel
PDCP	Packet Data Convergence Protocol
PDI	Packet Detection Information
PDN	Packet Data Network
PDP	Packet Data Protocol
PDR	Packet Detection Rule
PDSCH	Physical Downlink Shared Channel
PDU	Protocol Data Unit
PEI	Permanent Equipment Identifier
PFCP	Packet Forwarding Control Protocol
PGW	Packet Data Network Gateway
PIN	Personal IoT Network
PKI	Public Key Infrastructure
PLMN	Public Land Mobile Network
PNF	Physical Network Function
PON	Passive Optical Network
PPP	Public Private Partnership
PRACH	Physical Random Access Channel
ProSe	Proximity-based Services
PS	Packet Switched
PSA	PDU Session Anchor
PSTN	Public Switched Telephone Network
PT	Payload Type
PUCCH	Physical Uplink Control Channel
PUSCH	Physical Uplink Shared Channel

Q

QAM	Quadrature Amplitude Modulation
QER	QoS Enforcement Rule
QoE	Quality of Experience
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying

R

RADIUS	Remote Authentication Dial In User Service
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Resource Block
RCF	Radio Control Function
RedCap	Reduced Capability
REST	Representational State Transfer
RESTful	Representational State Transfer
RF	Radio Frequency
RFC	Request for Comments
RG	Residential Gateway
RIC	RAN Intelligent Controller
RIS	Reconfigurable Intelligent Surface
RLC	Radio Link Control
RNC	Radio Network Controller
RRC	Radio Resource Control
RRH	Remote Radio Head
RT	Real-Time
RTP	Real-time Transport Protocol
RTR	Rundfunk und Telekom Regulierungs-GmbH
RU	Radio Unit

S

SA	Standalone
SAA	Server-Assignment-Answer
SAE	System Architecture Evolution
SAR	Server-Assignment-Request
SAR	Specific Absorption Rate
SAS	Spectrum Access System
SBA	Service Based Architecture
SBC	Session Border Controller
SBC-M	Session Border Controller-Media
SBC-S	Session Border Controller-Signaling
SBI	South Bound Interface
SBI	Service-based Interface
SCC AS	Service Centralization and Continuity Application Server
SCN	Switched Circuit Network
SCP	Service Communication Proxy
S-CSCF	Serving-Call Session Control Function
SCTP	Stream Control Transmission Protocol

SD	Slice Differentiator
SDAP	Service Data Adaptation Protocol
SDG	Sustainable Development Goal
SDL	Supplementary Downlink
SDN	Software Defined Networking
SDP	Session Description Protocol
SEAF	SEcurity Anchor Function
SEPP	Security Edge Protection Proxy
SFC	Service Function Chaining
SFF	Service Function Forwarder
SFP	Service Function Path
SG	Study Group
SGSN	Serving GPRS Support Node
SGW	Serving Gateway
SGW	Signalling Gateway
S-GW	Serving-GW
SHD	Super HD Video
SI	Service Index
SIC	Self Interference Cancellation
SIDF	Subscription Identifier De-concealing Function
SIGTRAN	SIGnalling TRANsport
SloT	Socialized Internet of Things
SIP	Session Initiation Protocol
SL	Subscriber Locator
SLF	Subscription Locator Function
SL-IoT	Sidelink enhanced-Industry IoT
SM	Session Management
SM-MIMO	Spatial Modulation-MIMO
SMF	Session Management Function
SMO	Service Management and Orchestration
SMS	Short Message Service
SMSF	Short Message Service Function
SN	Serving Network
SNPN	Standalone Non-Public Network
SNS	Smart Networks and Services
SNS JU	Smart Networks and Services Joint Undertaking
S-NSSAI	Single-NSSAI
SON	Self-Organizing Networks
SPI	Service Path Identifier
SPN	Special Purpose Network
SRI	Satellite Radio Interface

SRVCC	Single Radio Voice Call Continuity
SSC	Session and Service Continuity
SST	Slice/Service Type
STIN	Space-terrestrial integrated network
STP	Signaling Transfer Point
SUCI	Subscription Concealed Identifier
SUL	Supplementary Uplink
SUPI	Subscription Permanent Identifier
SW	Software

T

TAS	Telephony Application Server
TCP	Transport Control Protocol
TDD	Time Division Duplexing
TDoS	Telephone Denial of Service
TIRO	Tactile Internet for remote operations
TKG	Telekommunikationsgesetz
TKK	Telekom-Control-Kommission
TLS	Transport Layer Security
TN	Transport Network
TNAN	Trusted Non-3GPP Access Network
TNAP	Trusted Non-3GPP Access Point
TNGF	Trusted Non-3GPP Gateway Function
TPM	Trusted Platform Module
TR	Technical Report
TrGW	Transition Gateway
TS	Technical Specification
TSDSI	Telecommunications Standards Development Society, India
TSN	Time-sensitive Networking
TTL	Time To Live
TUP	Telephone User Part
TWAP	Trusted WLAN Access Point
TWIF	Trusted WLAN Interworking Function

U

U	Unassigned
U	User data
UAA	User-Authorization-Answer
UAC	User Agent Client
UAM	Urban Air Mobility
UAR	User-Authorization-Request

UAS	Unmanned Aerial System
UAS	Unmanned Aircraft System
UAS	User Agent Server
UAV	Unmanned Aerial Vehicle
UDM	Unified Data Management
UDP	User Datagram Protocol
UDR	Unified Data Repository
UDSF	Unstructured Data Storage Function
UE	User Equipment
UEFI	Unified Extensible Firmware Interface
UHD	Ultra High Definition
UL	Up Link
umMTC	Ultra-Massive Machine-Type Communications
UMTS	Universal Mobile Telecommunications System
UN	United Nations
UP	User Plane
UPF	User Plane Function
URI	Uniform Resource Identifier
URLLC	Ultra-Reliable and Low Latency Communications
URR	Usage Reporting Rule
USIM	UMTS Subscriber Identity Module
USIM	Universal Subscriber Identity Module
UTRAN	Universal Terrestrial Radio Access Network
UVEK	Umwelt, Verkehr, Energie und Kommunikation

V

V	Volt
V2X	Vehicle-to-Everything
V2X	Vehicle to X
VANC	VoLGA Access Network Controller
VGCS	Voice Group Call Services
VIM	Virtualised Infrastructure Manager
VLAN	Virtual LAN
VLC	Visible Light Communication
VLR	Visitor Location Register
VLV&TIC	Very Large Volume & Tiny Instant Communications
VM	Virtual Machine
VNF	Virtualised Network Function
VNFM	VNF Manager
VoIP	Voice over IP
VoLGA	Voice over LTE over Generic Access Network

VoLTE	Voice over LTE
VoWifi	Voice over Wifi
VoWLAN	Voice over WLAN
VPLMN	Visited Public Land Mobile Network
VPN	Virtual Private Network
VR	Virtual Reality
vSEPP	visited SEPP
VXLAN	Virtual Extensible LAN

W

W	Watt
W-5GAN	Wireline 5G Access Network
W-5GBAN	Wireline 5G BBF Access Network
W-5GCAN	Wireline5G Cable Access Network
W-AGF	Wireline Access Gateway Function
W-CDMA	Wideband-Code Division Multiple Access
WebRTC	Web Real-Time Communication between Browsers
WHO	World Health Organization
Wifi	Wireless fidelity
WP	Working Party
WRC	World Radiocommunication Conference

X

XaaS	X as a Service
XCAP	XML Configuration Access Protocol
XMPP	Extensible Messaging and Presence Protocol
XnAP	Xn Application Protocol
XR	eXtended Reality

Z

ZVEI	Zentralverband Elektrotechnik- und Elektronikindustrie
------	--

References

- [1] Farinacci, D.; Li, T.; Hanks, S.; Meyer, D.; Traina, P.: RFC 2784 – Generic Routing Encapsulation (GRE). IETF, March 2000
- [2] Rosen, E.; Viswanathan, A.; Callon, R.: RFC 3031 – Multiprotocol Label Switching Architecture. IETF, January 2001
- [3] Rosenberg, J.; Schulzrinne, H.; Camarillo, G.; Johnston, A.; Peterson, J.; Sparks, R.; Handley, M.; Schooler, E.: RFC 3261 – SIP: Session Initiation Protocol. IETF, June 2002
- [4] Rosenberg, J.; Schulzrinne, H.: RFC 3264 – An Offer/Answer Model with the Session Description Protocol (SDP). IETF, June 2002
- [5] Schulzrinne, H.; Casner, S.; Frederick, R.; Jacobson, V.: RFC 3550 – RTP: A Transport Protocol for Real-Time Applications. IETF, July 2003
- [6] Schulzrinne, H.; Casner, S.: RFC 3551 – RTP Profile for Audio and Video Conferences with Minimal Control. IETF, July 2003
- [7] Loughney, J.: RFC 3589 – Diameter Command Codes for Third Generation Partnership Project (3GPP) Release 5. IETF, September 2003
- [8] Aboba, B.; Blunk, L.; Vollbrecht, J.; Carlson, J.; Levkowetz, H.: RFC 3748 – Extensible Authentication Protocol (EAP). IETF, June, 2004
- [9] Handley, M.; Jacobson, V.; Perkins, C.: RFC 4566 – SDP: Session Description Protocol. IETF, July 2006
- [10] Sterman, B.; Sadolevsky, D.; Schwartz, D.; Williams, D.; Beck, W.: RFC 5090 – RADIUS Extension for Digest Authentication. IETF, July 2006
- [11] Garcia-Martin, M.; Belinchon, M.; Pallares-Lopez, M.; Canales-Valenzuela, C.; Tammi, K.: RFC 4740 – Diameter Session Initiation Protocol (SIP) Application. IETF, November 2006
- [12] Fajardo, V.; Arkko, J.; Loughney, J.; Zorn, G.: RFC 6733 – Diameter Base Protocol. IETF, October 2012
- [13] Hardt, D.: RFC 6749 – The OAuth 2.0 Authorization Framework. IETF, October 2012
- [14] Mahalingam, M.; Dutt, D.; Duda, K.; Agarwal, P.; Kreeger, L.; Sridhar, T.; Bursell, M.; Wright, C.: RFC 7348 – Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks. IETF, August 2014
- [15] Quinn, P.; Nadeau, T.: RFC 7498 – Problem Statement for Service Function Chaining. IETF, April 2015
- [16] Garg, P.; Wang, Y.: RFC 7637 – NVGRE: Network Virtualization Using Generic Routing Encapsulation. IETF, September 2015
- [17] Halpern, J.; Pignataro, C.: RFC 7665 – Service Function Chaining (SFC) Architecture. IETF, 2015
- [18] Quinn, P.; Elzur, U.; Pignataro, C.: RFC 8300 – Network Service Header (NSH). IETF, January 2018
- [19] TR 21.915 V15.0.0: Release 15 Description; Summary of Rel-15 Work Items (Release 15). 3GPP, September 2019

- [20] TR 21.916 V16.2.0: Release 16 Description; Summary of Rel-16 Work Items (Release 16). 3GPP, June 2022
- [21] TS 22.261 V15.8.0: Service requirements for the 5G system; Stage 1 (Release 15). 3GPP, September 2019
- [22] TS 22.261 V16.10.0: Service requirements for the 5G system; Stage 1 (Release 16). 3GPP, December 2019
- [23] IEEE Std C95.1-2019: IEEE Standard for Safety Levels with Respect to Human Exposure to Electric, Magnetic, and Electromagnetic Fields, 0 Hz to 300 GHz. IEEE Standards Coordinating Committee 39, February 2019
- [24] TR 22.804 V16.2.0: Study on Communication for Automation in Vertical Domains (Release 16). 3GPP, December 2018
- [25] TR 22.822 V16.0.0: Study on using Satellite Access in 5G; Stage 1 (Release 16). 3GPP, June 2018
- [26] TR 22.886 V15.3.0: Study on enhancement of 3GPP Support for 5G V2X Services (Release 15). 3GPP, September 2018
- [27] TR 22.891 V14.2.0: Feasibility Study on New Services and Markets Technology Enablers; Stage 1 (Release 14). 3GPP, September 2016
- [28] TS 23.002 V4.8.0: Network architecture (Release 4). 3GPP, June 2003
- [29] TS 23.002 V5.12.0: Network architecture (Release 5). 3GPP, September 2003
- [30] TS 23.002 V8.7.0: Network architecture (Release 8). 3GPP, December 2010
- [31] TS 23.002 V3.1.0: Network architecture (Release 99). 3GPP, September 1999
- [32] TS 23.216 V12.2.0: Single Radio Voice Call Continuity (SRVCC); Stage 2 (Release 12). 3GPP, December 2014
- [33] TS 23.228 V15.4.0: IP Multimedia Subsystem (IMS); Stage 2 (Release 15). 3GPP, March 2019
- [34] TS 23.237 V10.13.0: IP Multimedia Subsystem (IMS) Service Continuity; Stage 2 (Release 10). 3GPP, December 2015
- [35] TS 23.272 V12.4.0: Circuit Switched (CS) fallback in Evolved Packet System (EPS); Stage 2 (Release 12). 3GPP, September 2014
- [36] TS 23.402 V8.10.0: Architecture enhancements for non-3GPP accesses (Release 8). 3GPP, March 2012
- [37] TS 23.501 V16.14.0: System Architecture for the 5G System (5GS); Stage 2 (Release 16). 3GPP, September 2022
- [38] TS 23.501 V16.2.0: System Architecture for the 5G System (5GS); Stage 2 (Release 16). December 2022
- [39] TS 23.502 V16.15.0: Procedures for the 5G System (5GS); Stage 2 (Release 16). 3GPP, December 2022
- [40] TR 23.799: Study on Architecture for Next Generation System (Release 14). 3GPP, December 2016

- [41] TS 24.228 V5.15.0: Signalling flows for the IP multimedia call control based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3 (Release 5). 3GPP, September 2006
- [42] TS 28.530 V16.5.0: Management and Orchestration; Concepts, use cases and requirements (Release 16). 3GPP, December 2021
- [43] TS 29.500 V16.12.0: 5G System; Technical Realization of Service Based Architecture; Stage 3 (Release 16). 3GPP, September 2022
- [44] TS 29.501 V16.7.0: 5G System; Principles and Guidelines for Services Definition; Stage 3 (Release 16). 3GPP, September 2022
- [45] TS 33.501 V16.13.0: Security architecture and procedures for 5G system (Release 16). 3GPP, December 2022
- [46] TS 37.340 V15.6.0: Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Multi-connectivity; Stage 2 (Release 15). 3GPP, June 2019
- [47] TS 38.104 V15.7.0: NR; Base Station (BS) radio transmission and reception (Release 15). 3GPP, September 2019
- [48] TS 38.201 V15.0.0: NR; Physical layer; General description (Release 15). 3GPP, December 2017
- [49] TS 38.202 V15.6.0: NR; Services provided by the physical layer (Release 15). 3GPP, December 2019
- [50] TS 38.211 V1.0.0: NR; Physical channels and modulation (Release 15). 3GPP, September 2017
- [51] TS 38.300 V16.0.0: NR; NR and NG-RAN Overall Description; Stage 2 (Release 15). 3GPP, September 2022
- [52] TR 38.821 V16.0.0: Solutions for NR to support non-terrestrial networks (NTN) (Release 16). 3GPP, December 2019
- [53] 26. BImSchV: Sechsundzwanzigste Verordnung zur Durchführung des Bundes-Immissionsschutzgesetzes (Verordnung über elektromagnetische Felder - 26. BImSchV). Bundesrepublik Deutschland, 16.12.1996, neugefasst am 14.8.2013
- [54] <https://www.3gpp.org/about-3gpp>
- [55] <https://www.3gpp.org/3gpp-groups>
- [56] TR 37.910 V16.1.0: Study on self evaluation towards IMT-2020 submission (Release 16). 3GPP, September 2019
- [57] <https://www.3gpp.org/specifications-technologies/releases>
- [58] <https://www.3gpp.org/specifications-technologies/releases/release-18>
- [59] <https://www.3gpp.org/specifications-technologies/releases/release-15>
- [60] <https://www.3gpp.org/specifications-technologies/releases/release-16>
- [61] <https://www.3gpp.org/specifications-technologies/releases/release-17>
- [62] TR 37.910 V17.0.0: Study on self evaluation towards IMT-2020 submission (Release 17). 3GPP, March 2022
- [63] <https://www.3gpp.org/DynaReport/FeatureListFrameSet.htm>
- [64] <https://5gaa.org>

- [65] 5GAA: C-V2X Use Cases – Methodology, Examples and Service Level Requirements, Whitepaper. 5GAA, June 2019
- [66] <https://5g-ppp.eu>
- [67] <http://www.6gflagship.com>
- [68] <http://www.6gsummit.com>
- [69] IEEE Std 802.1Q-2014: Bridges and Bridged Networks. IEEE, November 2014
- [70] <https://www.5g-acia.org>
- [71] 5G-ACIA: 5G for Connected Industries and Automation, Whitepaper, 2nd edition. 5G-ACIA, February 2019
- [72] Agouros, Konstantinos: Software Defined Networking: SDN-Praxis mit Controllern und OpenFlow. De Gruyter, 2017
- [73] Badach, Anatol: Protokolle und Dienste der Informationstechnologie – SFC Service Function Chaining. WEKA, Januar 2015
- [74] <https://www.bakom.admin.ch>
- [75] <https://www.bakom.admin.ch/bakom/de/home/frequenzen-antennen/vergabe-der-mobilfunkfrequenzen/mobilfunkfrequenzen-5G-vergeben.html>
- [76] BAKOM: Ausschreibung von Frequenzblöcken für die landesweite Erbringung von mobilen Fernmeldediensten in der Schweiz. BAKOM, Juli 2018
- [77] Balapuwaduge, Indika A. M.; Li, Frank Y.: Cellular Networks: An Evolution from 1G to 4G. Aus: Encyclopedia of Wireless Networks. Springer, 2020
- [78] <https://www.bfs.de>
- [79] BFS: 5G – the fifth generation of mobile communications: Standpoint on Radiation Protection. BFS, May 2021
- [80] <https://www.bfs.de/DE/themen/emf/mobilfunk/vorsorge/recht/grenzwerte.html>
- [81] <https://www.bundesnetzagentur.de>
- [82] <https://www.bundesnetzagentur.de/DE/Fachthemen/Telekommunikation/Breitband/MobilesBreitband/Frequenzauktion/2019/Auktion2019-node.html>
- [83] BNetzA: Entscheidung der Präsidentenkammer der Bundesnetzagentur für Elektrizität, Gas, Telekommunikation, Post und Eisenbahnen vom 26. November 2018 über die Festlegungen und Regeln im Einzelnen (Vergaberegeln) und über die Festlegungen und Regelungen für die Durchführung des Verfahrens (Auktionsregeln) zur Vergabe von Frequenzen in den Bereichen 2 GHz und 3,6 GHz, Aktenzeichen: BK1- 17/001. BNetzA, November 2018
- [84] Bundesnetzagentur: Katalog von Sicherheitsanforderungen für das Betreiben von Telekommunikations- und Datenverarbeitungssystemen sowie für die Verarbeitung personenbezogener Daten nach § 109 Telekommunikationsgesetz (TKG) Version 2.0. Bundesnetzagentur, 29.4.2020
- [85] <https://www.broadband-forum.org>
- [86] Broadband Forum: MR-464 - Migrating Fixed Access to 5G Core, Issue: 1. Broadband Forum, October 2019
- [87] <https://www.cablelabs.com>

- [88] Chandramouli, Devaki; Liebhart, Rainer; Pirskanen, Juho: 5G for the Connected World. Wiley, 2019
- [89] Chayapathi, Rajendra; Hassan, Syed Farrukh; Shah, Paresh: Network Functions Virtualization (NFV) with a Touch of SDN. Addison-Wesley, 2017
- [90] Chiosi, M. et al.: Network Functions Virtualization – An Introduction, Benefits, Enablers, Challenges & Call for Action. ETSI Whitepaper, October 2012
- [91] Cox, Christopher: An Introduction to LTE – LTE, LTE-Advanced, SAE, VoLTE and 4G Mobile Communications. Wiley, 2014
- [92] Delb, Valentin; Dudle, Gregor; Dürrenberger, Gregor; Grasser, Christian; Horisberger, Philippe; Künzle, Harry; Kuster, Niels; Netzle, Stephan; Portmann, Manfred; Quinto, Carlos; Reichenbach, Alexander; Rösli, Martin; Siegenthaler, Andreas; Steffen, Paul; Steiner, Edith; Stempfel, Evelyn; Stijve, Sanne; Studerus, Jürg; Walker, Urs; Weber, Felix; Ziebold, Rolf: Bericht Mobilfunk und Strahlung. Arbeitsgruppe Mobilfunk und Strahlung im Auftrag des UVEK, November 2019
- [93] Eckert, Claudia: IT-Sicherheit – Konzepte - Verfahren - Protokolle. De Gruyter, 2023
- [94] Delb, Valentin; Dudle, Gregor; Dürrenberger, Gregor; Grasser, Christian; Horisberger, Philippe; Künzle, Harry; Kuster, Niels; Netzle, Stephan; Portmann, Manfred; Quinto, Carlos; Reichenbach, Alexander; Rösli, Martin; Siegenthaler, Andreas; Steffen, Paul; Steiner, Edith; Stempfel, Evelyn; Stijve, Sanne; Studerus, Jürg; Walker, Urs; Weber, Felix; Ziebold, Rolf: Report Mobile Radio and Radiation. Working group on Mobile Radio and Radiation on behalf of the Federal Department of the Environment, Transport, Energy and Communications (DETEC), November 2019
- [95] Giust, Fabio et al.: MEC Deployments in 4G and Evolution Towards 5G. ETSI Whitepaper, February 2018
- [96] El Hattachi, R.; Erfanian, J.: NGMN 5G Initiative 5G Whitepaper, Version 1.0. NGMN, February 2015
- [97] GS NFV-EVE 005 V1.1.1: Report on SDN Usage in NFV Architectural Framework. ETSI, December 2015
- [98] Fallgren, Mikael et al.: Scenarios, requirements and KPIs for 5G mobile and wireless system, D1.1. EU Project METIS, April 2013
- [99] <https://www.itu.int/en/ITU-T/focusgroups/net2030>
- [100] FG Cloud TR Version 1.0: Part 2: Functional requirements and reference architecture. ITU-T, Focus Group on Cloud Computing, February 2012
- [101] Frenger, Pal; Tano, Richard: More Capacity and Less Power: How 5G NR can Reduce Network Energy Consumption. 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring), Kuala Lumpur, Malaysia, 2019, pp. 1-5
- [102] Göransson, Paul; Black, Chuck: Software Defined Networks – A Comprehensive Approach. Elsevier, 2014
- [103] Guttmann, Erik; Ali, Irfan: Path to 5G: A Control Plane Perspective. Journal of ICT, 2018, Vol. 6, No. 1-2, pp. 87-100
- [104] H.248.1: Gateway control protocol: Version 3. ITU-T, March 2013

- [105] Höfer, Tim; Bierwirth, Sebastian; Madlener, Reinhard: Energie-Mehrverbrauch in Rechenzentren bei Einführung des 5G Standards. Studie. FCN RWTH Aachen, August 2019
- [106] Holma, Harri; Toskala, Antti; Nakamura, Takehiro: 5G Technology – 3GPP New Radio. Wiley, 2020
- [107] Huang, T.; Yang, W.; Wu, J.; Ma, J.; Zhang, X.; Zhang, D.: A Survey on Green 6G Network: Architecture and Technologies. *IEEE Access*, vol. 7, pp. 175758-175768, December 2019
- [108] Huawei: 5G Power Whitepaper. Huawei, February 2019
- [109] I, Chih-Lin; Katti, Sachin: O-RAN: Towards an Open and Smart RAN – Whitepaper. O-RAN Alliance, October 2018
- [110] <https://www.icnirp.org>
- [111] ICNIRP: ICNIRP Guidelines – For Limiting Exposure to Time-Varying Electric, Magnetic and Electromagnetic Fields (up to 300 GHz). Published in: *Health Physics* 74 (4), pp 494-522, 1998
- [112] ICNIRP: ICNIRP Guidelines – For Limiting Exposure to Electromagnetic Fields (100 kHz to 300 GHz). Published in: *Health Physics* 118(5): 483–524; 2020
- [113] <https://www.icnirp.org/en/differences.html>
- [114] IR.92: IMS Profile for Voice and SMS, Version 7.0. GSMA, March 2013
- [115] <https://www.itu.int/en/ITU-R>
- [116] Jayakody, Dushantha Nalin K.; Srinivasan, Kathiravan; Sharma, Vishal: 5G Enabled Secure Wireless Networks. Springer, 2019
- [117] Johnston, Alan B.: SIP – Understanding the Session Initiation Protocol. 3. Ed., Artech House, 2015
- [118] Kappes, Martin: Netzwerk- und Datensicherheit. Springer Vieweg, 2023
- [119] M.2150-1: Detailed specifications of the terrestrial radio interfaces of International Mobile Telecommunications-2020 (IMT-2020). ITU-R, February 2022
- [120] Keith, Robert: 5G Energy Efficiency Explained. A10 Networks, 10.3.2020
- [121] Kühn, Paul: Vorlesungsskript Nachrichtenvermittlung I und II. Universität Stuttgart, Institut für Nachrichtenvermittlung und Datenverarbeitung, 1991
- [122] Lédl, Petr et al.: O-RAN Town: Piloting a High-Power Multivendor Open RAN Solution in a Brownfield Network. Whitepaper. Deutsche Telekom, 2023
- [123] Li, Richard et al.: Network 2030 – A Blueprint of Technology, Applications and Market Drivers Towards the Year 2030 and Beyond. Whitepaper. ITU-T FG-NET2030, May 2019
- [124] Li, Richard: Network 2030 and New IP. IEEE CNSM 2019, Halifax/Canada, 21.-25. October 2019
- [125] Liyanage, Madhusanka; Ahmad, Ijaz; Abro, Ahmed Bux; Gurtov, Andrei; Ylianttila, Mika: A Comprehensive Guide to 5G Security. Wiley, 2018
- [126] Lourenco, Marco; Marinos, Louis: ENISA Threat Landscape for 5G Networks. ENISA, November 2019
- [127] M.2376-0: Technical feasibility of IMT in bands above 6 GHz. ITU-R, July 2015
- [128] M.2083-0: IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond. ITU-R, September 2015

- [129] MacKenzie, Richard: NGMN Overview on 5G RAN Functional Decomposition. NGMN Alliance, Feb. 2018
- [130] Mademann, Frank: The 5G System Architecture. Journal of ICT, 2018, Vol. 6, No. 1-2, p. 77-86
- [131] Malik, Sukhwinder Singh; Atri, Rahul: 5G New Radio – Next Generation of Mobile Broadband: 5G New Radio Technology Introduction and its Throughput Capabilities. Whitepaper. 24.6.2018
- [132] GS NFV-MAN 001 V1.1.1: Network Functions Virtualisation (NFV); Management and Orchestration. ETSI, December 2014
- [133] Marsch, Patrick; Bulakci, Ömer; Queseth, Olav; Boldi, Mauro: 5G System Design – Architectural and Functional Considerations and Long Term Research. Wiley, 2018
- [134] Maternia, Michal et al.: 5G PPP use cases and performance evaluation models, Version 1.0. 5GPPP, April 2016
- [135] Mayer, Georg: RESTful APIs for the 5G Service Based Architecture. Journal of ICT, 2018, Vol. 6, No. 1-2, p. 101-116
- [136] GS MEC 002 V2.1.1: Multi-access Edge Computing (MEC); Phase 2: Use Cases and Requirements. ETSI, October 2018
- [137] GS MEC 003 V2.1.1: Multi-access Edge Computing (MEC); Framework and Reference Architecture. ETSI, January 2019
- [138] GS MEC-IEG 004 V1.1.1: Mobile-Edge Computing (MEC); Service Scenarios. ETSI, November 2015
- [139] <https://metis2020.com>
- [140] GR NFV 001 V1.2.1: Network Functions Virtualisation (NFV); Use Cases. ETSI, May 2017
- [141] GS NFV 002 V1.2.1: Network Functions Virtualisation (NFV); Architectural Framework. ETSI, December 2014
- [142] Schweizerischer Bundesrat: 814.710 – Verordnung über den Schutz vor nichtionisierender Strahlung (NISV). 23. Dezember 1999 (Stand am 1. Juni 2019)
- [143] <https://www.ngmn.org>
- [144] Nokia: 5G network energy efficiency. Whitepaper. Nokia, 2016
- [145] Nokia: Zero Emissions Networks - Turning the zero carbon vision into business opportunity. Nokia, 2019
- [146] ITU-R WP 5D: IMT.Framework for 2030 and Beyond, Draft. ITU-R, June 2023
- [147] ONF: OpenFlow Switch Specification, Version 1.0.0. ONF, December 2009
- [148] <https://www.opennetworking.org>
- [149] ONF: OpenFlow Switch Specification, Version 1.5.1. ONF, March 2015
- [150] <https://www.o-ran.org>
- [151] Osseiran, Afif et al.: Scenarios for the 5G Mobile and Wireless Communications: the Vision of the METIS Project. IEEE Communications Magazine Vol. 52 Issue 5 pp.26-35, May 2014
- [152] One6G: Taking communications to the next level. Position Paper. One6G, November 2021

- [153] Penttinen, Jyrki T. J.: 5G Explained - Security and Deployment of Advanced Mobile Communications. Wiley, 2019
- [154] Poikselkä, Miikka; Mayer, Georg: The IMS – IP Multimedia Concepts and Services. John Wiley, 2009
- [155] Pujol, Frédéric; Manero, Carole; Remis, Santiago: 5G Observatory Quarterly Report 5 – Up to September 2019, Study for European Commission. IDATE DIGIWORLD, October 2019
- [156] Rijsman, B.; Moisand, J.: draft-rijsman-sfc-metadata-considerations-00.txt – Metadata Considerations. IETF, February 2014
<https://www.rtr.at>
- [157] Sauter, Martin: From GSM to LTE-Advanced Pro and 5G. Wiley, 2017
- [158] Schneider, Peter; Urban, Josef: Die Sicherheitsarchitektur von Mobilfunknetzen. ITG News, 1/2020, S. 4-7
- [160] GS NFV-SEC 003 V1.1.1: NFV Security; Security and Trust Guidance. ETSI, December 2014
- [161] Siegmund, Gerd: Technik der Netze 1. VDE-Verlag, 2014
- [162] Siegmund, Gerd: Technik der Netze 2. VDE-Verlag, 2014
- [163] Siegmund, Gerd: SDN: Software-defined Networking – Neue Anforderungen und Netzarchitekturen für performante Netze. VDE-Verlag, 2018
- [164] Simko, Myrtill; Mattsson, Mats-Olof: 5G Wireless Communication and Health Effects – A Pragmatic Review Based on Available Studies Regarding 6 to 100 GHz. Int. Journal of Environmental Research and Public Health, 2019
- [165] Stallings, William: Data and Computer Communications. Pearson, 2014
- [166] Stallings, William: Foundations of Modern Networking – SDN, NFV, QoE, IoT, and Cloud. Pearson, 2016
- [167] Stobbe, Lutz; Kemkes, Michael; Mager, Thomas; Oberthür, Simon; Schomaker, Gunnar: Whitepaper – 5G Charakterisierung, Version 1.1. Paderborn, 2019
- [168] FG NET-2030 Sub-G2: New Services and Capabilities for Network 2030: Description, Technical Gap and Performance Target Analysis. ITU-T FG NET-2030, October 2019
- [169] Teral, Stephane: 5G best choice architecture – Whitepaper. IHS Markit, 2019
- [170] <https://telecomprotocols.blogspot.com/2012/09/h248megaco-protocol.html>
- [171] TKK: Bescheid F 7/16-401. Telekom-Control-Kommission, April 2019
- [172] TKK: Anlage zum Bescheid F 7/16-401 der Telekom-Control-Kommission vom 08.04.2019. Telekom-Control-Kommission, April 2019
- [173] Trick, Ulrich; Weber, Frank: SIP und Telekommunikationsnetze. 5. Auflage. De Gruyter Oldenbourg, 2015
- [174] Vaezi, Mojtaba; Zhang, Ying: Cloud Mobile Networks – From RAN to EPC. Springer, 2017
- [175] Westerberg, Eric: 4G/5G RAN architecture – how a split can make the difference. Ericsson Technology Review, July 2016
- [176] <https://news.itu.int/wrc-19-agrees-to-identify-new-frequency-bands-for-5g>

- [177] ITU-R: World Radiocommunication Conference 2019 (WRC-19) – Provisional Final Acts. ITU-R, October – November 2019
- [178] Y.2001: General Overview of NGN. ITU-T, December 2004
- [179] Y.3001: Future networks: Objectives and design goals. ITU-T, May 2011
- [180] Y.3011: Framework of network virtualization for future networks. ITU-T, January 2012
- [181] Y.3300: Framework of software-defined networking. ITU-T, June 2014
- [182] Y.3502: Information technology – Cloud computing – Reference architecture. ITU-T, August 2014
- [183] Next G Alliance: 6G Technologies, Report. Next G Alliance, June 2022
- [184] Zhu, Shao Ying; Scott-Hayward, Sandra; Jacquin, Ludovic; Hill, Richard: Guide to Security in SDN and NFV - Challenges, Opportunities, and Applications. Springer, 2017
- [185] TS 22.261 V16.16.0: Service requirements for the 5G system; Stage 1 (Release 16). 3GPP, December 2021
- [186] TS 22.261 V17.11.0: Service requirements for the 5G system; Stage 1 (Release 17). 3GPP, September 2022
- [187] TS 38.104 V16.13.0: NR; Base Station (BS) radio transmission and reception (Release 16). 3GPP, September 2022
- [188] TS 38.104 V17.7.0: NR; Base Station (BS) radio transmission and reception (Release 17). 3GPP, September 2022
- [189] Wikström, Gustav et al.: What societal values will 6G address? Societal Key Values and Key Value Indicators analysed through 6G use cases. 6G IA, May 2022
- [190] One6G: 6G Technology Overview. White Paper, 2nd Ed. One6G, November 2022
- [191] Jiang, Wie; Han, Bin; Habibi, Mohammad Asif; Schotten, Hans Dieter: The Road Towards 6G: A Comprehensive Survey. IEEE Open Journal of the Communications Society, vol. 2, pp. 334-366, 2021
- [192] Samsung: 6G – The Next Hyper-Connected Experience for All, Whitepaper. Samsung, 2020
- [193] <https://www.fcc.gov/auctions>
- [194] <https://www.fcc.gov>
- [195] Huawei: 6G: The Next Horizon, Whitepaper. Huawei, 2021
- [196] 5G Americas: 5G Spectrum Vision. Whitepaper. 5G Americas, February 2019
- [197] <https://5gobservatory.eu/national-5g-spectrum-assignment>
- [198] ESPI: C-Band, Satellites and 5G. ESPI Briefs No. 42. European Space Policy Institute, June 2020
- [199] <https://www.fcc.gov/wireless/bureau-divisions/mobility-division/35-ghz-band/35-ghz-band-overview>
- [200] FCC: 2020 Communications Marketplace Report. FCC-188, December 2020
- [201] IMT 2030 (6G) Promotion Group: Views towards IMT for 2030 and Beyond. IMT 2030 (6G) Promotion Group, January 2023
- [202] Next G Alliance: 6G Applications and Use Cases. Next G Alliance, 2022

- [203] Hoffmann, Marco; Uusitalo, Mikko; Hamon, Marie-Helene; Richerzhagen, Björn; D`Aria, Giovanna; Gati, Azedinne; Lopez, Diego et al.: Expanded 6G vision, use cases and societal values. Deliverable D1.2. Hexa-X, April 2021
- [204] Zhao, Quan; Lister, David; Saxena, Narothum: 6G Use Cases and Analysis, Version 1.0. NGMN, February 2022
- [205] Taleb, Tarik et al.: White Paper on 6G Networking. 6G Flagship, University of Oulu, June 2020
- [206] VVA; Policytracker; LS: 5G Observatory Biannual Report – April 2023, Study on European 5G Observatory phase III. CNECT/2021/OP/0008, April 2023
<https://www.fcc.gov/auction/1000>
- [208] Qualcom: Global update on spectrum for 4G & 5G. December 2020
- [209] O-RAN WG1: O-RAN Architecture Description 7.0. O-RAN Alliance, October 2022
- [210] O-RAN WG1: O-RAN Use Cases Detailed Specification 9.0. O-RAN Alliance, October 2022
- [211] O-RAN WG1: O-RAN Use Cases Analysis Report 9.0. O-RAN Alliance, October 2022
- [212] O-RAN WG1: O-RAN Slicing Architecture 8.0. O-RAN Alliance, April 2020
- [213] TS 29.244 V16.10.0: Interface between the Control Plane and the User Plane Nodes; Stage 3 (Release 16). 3GPP, June 2022
- [214] Stewart, R.: RFC 4960 – Stream Control Transmission Protocol. IETF, September 2007
- [215] TS 38.410 V16.4.0: NG-RAN; NG general aspects and principles (Release 16). 3GPP, October 2021
- [216] TS 38.413 V16.11.0: NG-RAN; NG Application Protocol (NGAP) (Release 16). 3GPP, September 2022
- [217] TS 29.281 V16.2.0: General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U) (Release 16). 3GPP, March 2021
- [218] TS 38.300 V17.3.0: NR; NR and NG-RAN Overall Description; Stage 2 (Release 17). 3GPP, December 2022
- [219] TS 23.501 V17.7.0: System Architecture for the 5G System (5GS); Stage 2 (Release 17). 3GPP, December 2022
- [220] TS 28.531 V17.6.0: Management and orchestration; Provisioning; (Release 17). 3GPP, December 2022
- [221] NG.116: Generic Network Slice Template, Version 7.0. GSMA, June 2022
- [222] Y.2060: Overview of the Internet of things. ITU-T, June 2012
- [223] Zaidi, Ali; Bränneby, Anders; Nazari, Ala; Hogan, Marie; Kuhlins, Christian: Cellular IoT in the 5G era. Whitepaper. Ericsson, February 2020
- [224] 5G-ACIA: Integration of 5G with Time-Sensitive Networking for Industrial Communications, Whitepaper. 5G-ACIA, February 2021
- [225] BMWi: Leitfaden 5G-Campusnetze – Orientierungshilfe für kleine und mittelständische Unternehmen. Bundesministerium für Wirtschaft und Energie, Juni 2020
- [226] NG.123: 5G industry campus network deployment guideline, Version 1.0. GSMA, November 2020
- [227] BSI TR-03163: Sicherheit in TK-Infrastrukturen. BSI, 25.5.2022

- [228] Köpsell, Stefan; Ruzhanskiy, Andrey; Hecker, Andreas; Stachorra, Dirk: Open-RAN Risikoanalyse (5GRANR). BSI, 21.2.2022
- [229] Matinmikko-Blue, Marja et al.: Whitepaper on 6G Drivers and the UN SDGs. 6G Flagship, University of Oulu, June 2020
- [230] Ericsson: On the road to breaking the energy curve. Whitepaper. Ericsson, 2022
- [231] Huawei: Green 5G. Whitepaper. Huawei, 2021
- [232] Huawei: Green 5G: Building a Sustainable World. Whitepaper. Huawei, 2020
- [233] Brundtland, Gro Harlem et al.: Report of the World Commission on Environment and Development: Our Common Future. UN World Commission, March 1987
- [234] UN Resolution adopted by the General Assembly: Transforming our world: the 2030 Agenda for Sustainable Development. UN, September 2015
- [235] Pacchierotti, Francesca et al.: Effects of Radiofrequency Electromagnetic Field (RF-EMF) exposure on male fertility and pregnancy and birth outcomes: Protocols for a systematic review of experimental studies in non-human mammals and in human sperm exposed in vitro. Environment International Journal, Elsevier, August 2021
- [236] Belpoggi, Fiorella et al.: Health impact of 5G – Current state of knowledge of 5G-related carcinogenic and reproductive/developmental hazards as they emerge from epidemiological studies and in vivo experimental studies. European Parliamentary Research Service, July 2021
- [237] 5G Americas: Becoming 5G-Advanced: the 3GPP 2025 Roadmap. Whitepaper. 5G Americas, December 2022
- [238] TS 22.261 V18.8.0: Service requirements for the 5G system; Stage 1 (Release 18). 3GPP, December 2022
- [239] TS 38.104 V18.0.0: NR; Base Station (BS) radio transmission and reception (Release 18). 3GPP, December 2022
- [240] FG NET-2030 Sub-G1 Technical Report: Representative Use Cases and Key Network Requirements for Network 2030. ITU-T FG NET-2030, January 2020
- [241] FG NET-2030 Sub-G1 Technical Report: Additional Representative Use Cases and Key Network Requirements for Network 2030. ITU-T FG NET-2030, June 2020
- [242] FG NET-2030 Sub-G2 Technical Specification: Network 2030 Architecture Framework. ITU-T FG NET-2030, June 2020
- [243] TR 21.917 V17.0.1: Release 17 Description; Summary of Rel-17 Work Items (Release 17). 3GPP, January 2023
- [244] <https://www.itu.int/en/itu-r/study-groups/rsg5/rwp5d/Pages/default.aspx>
- [245] Bernardos, Carlos J.; Uusitalo, Mikko A. et al.: European Vision for the 6G Network Ecosystem, Version 1.0. 5G IA, June 2021
- [246] Ericson, Mårten; Flinck, Hannu; Vlacheas, Panagiotis; Wänstedt, Stefan et al.: Initial 6G Architectural Components and Enablers, Deliverable D5.1. Hexa-X, February 2022
- [247] <https://smart-networks.europa.eu>
- [248] <https://6g-ia.eu>
- [249] <https://hexa-x.eu>

- [250] <https://hexa-x-ii.eu>
- [251] <https://www.6gem.de>
- [252] <https://6g-life.de>
- [253] <https://6g-ric.de>
- [254] <https://www.open6ghub.de>
- [255] <https://www.6g-platform.com>
- [256] <https://www.nextgalliance.org>
- [257] B5GPC: Beyond 5G White Paper – Message to the 2030s. Version 1.51. Beyond 5G Promotion Consortium, October 2022
- [258] <https://www.imt2030.org.cn>
- [259] <https://one6g.org>
- [260] M.2516-0: Future technology trends of terrestrial International Mobile Telecommunications systems towards 2030 and beyond. ITU-R, November 2022
- [261] ITU-R WP 5D: Meeting Report #41. ITU-R, June 2022
- [262] <https://www.itu.int/wrc-23>
- [263] <https://www.surrey.ac.uk/institute-communication-systems/5g-6g-innovation-centre>
- [264] 5G Americas: Mobile Communications Towards 2030, White Paper. 5G Americas, November 2021
- [265] <https://tsdsi.in>
- [266] Samsung: 6G Spectrum – Expanding the Frontier, Whitepaper. Samsung, 2022
- [267] Kunzmann, Gerald: The 6G Future: Delivering new Levels of Customization, Resilience, and Privacy. Nokia, 27. VDE/ITG Fachtagung Mobilkommunikation, Osnabrück, May 2023
- [268] Chun, SungDuck: 3GPP Rel-19 Toward 6G. ofinno, September 2022
- [269] <https://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2030/Pages/default.aspx>
- [270] https://tiaonline.org/website_categories/5g
- [271] Lescuyer, Pierre; Lucidarme, Thierry: EvolvedPacket System (EPS) – The LTE and SAE Evolution of 3G UMTS. Wiley, 2008
- [272] <https://www.5gamerica.org>
- [273] <http://113.209.136.71/en/category/65573>
- [274] <http://www.5gforum.org/html/en/main.php>
- [275] <https://5gmf.jp/en>

Index

- 3GPP
 - Access 211
 - Networks 59
 - Organization 125
 - Release 10 204
 - Release 14 207
 - Release 15 115, 126, 135
 - Release 16 117, 126, 139
 - Release 17 119, 126, 140, 270
 - Release 18 267
 - Release 19 270
 - Release 4 13
 - Release 5 17, 30
 - Release 8 58
 - Release 99 10
 - Releases 30
 - Security architecture 242
 - Working groups 125
- 5G
 - Applications 107
 - Architecture 226
 - Core network 137f., 168, 178, 226
 - Design principles 134, 161, 167f.
 - Features 120, 135, 268
 - Network architecture 142, 226, 258
 - Overall view 229
 - System architecture 169, 178, 183
 - System overview 133, 182
 - Usage scenarios 104
- 5G PPP 281
- 5G system 114, 133, 135, 168, 182, 204, 226, 233
 - Phase 1 135
 - Phase 2 139
- 5GAA 109
- 5G-ACIA 108
- 5GC *See* 5G -Core network
- 5G-IA 279
- 6G
 - Architecture principles 301
 - Coverage 287, 292, 296
 - Flagship 281
 - Frequency ranges 283
 - Goals 279
 - Hyper Reliable and Low-Latency Communication 290
 - Immersive Communication 290
 - Integrated AI and Communication 291
 - Integrated Sensing and Communication 287, 291
 - Massive Communication 290
 - Network architecture 302
 - Orchestration 303
 - Overall view 304
 - RAN 297, 300, 305
 - Reasons 279f.
 - Requirements 292, 295
 - Research hubs 281
 - Security 280, 292, 298
 - Sustainability 279, 290, 292
 - System 304
 - Technologies 296
 - Ubiquitous Connectivity 292
 - Use cases 285, 289
 - Wireless Summit 278
- 6G SNS IA 281
- 6G-IA 281
- AAA 46, 236
- Absorption behavior 253
- Access Gateway Function *See* AGF
- Access Network 11, 30, 145, 161, 207, 211, 214, 226
- Access Stratum *See* AS
- Access Traffic Steering, Switching and Splitting *See* ATSSS
- A-CPI 89
- Advanced Persistent Threat *See* APT
- AF 180
- AGF 214
- AI 236, 272f., 275
- AI as a Service *See* AlaaS
- AlaaS 288
- AKA 244
- All over IP 204
- AMF 170, 207, 247
- Antenna 149, 154
 - Port 149
- API 76, 183
 - Northbound 89

- RESTful 90
- Southbound 83
- Application Layer Gateway 18, 235
- Application Plane 238
- Application Programming Interface *See API*
- Application Server 36
- Applications-Controller Plane Interface *See A-CPI*
- APT 232, 241
- AR 286
- AS 170
- ATM 9
- ATSSS 216
- Attribute-Value Pair *See AVP*
- AuC 11
- Augmented Reality *See AR*
- AUSF 180, 246
- Authentication 236, 244
 - Secondary 248
- Authentication, Authorization and Accounting *See AAA*
- Authorization 244
- Automation 67, 117, 228, 272
- AVP 46
- B5GPC 282
- Backhaul 226
- Backscattering 298
- Back-to-Back User Agent 18
- BAN 289, 304
- Base Band Unit *See BBU*
- BBU 72
- Beam
 - Forming 155, 249, 257, 260
 - Sweeping 156
- Beyond 5G Promotion Consortium *See B5GPC*
- BFS 254
- BGCF 34
- BICC 13
- Bidding down 243
- BNG 214
- Body Area Network *See BAN*
- Botnet 232, 234
- Breakout Gateway Control Function *See BGCF*
- BRG 215
- Broadband Forum 213
- Broadband Network Gateway *See BNG*
- Broadband Residential Gateway *See BRG*
- BSI 235
- BSS 228
- Bundesamt für Sicherheit in der Informationstechnik *See BSI*
- Bundesamt für Strahlenschutz *See BFS*
- Burst 278
 - Switching 278
- Burstlet 278
- CaaS 304
- Cable Residential Gateway *See CRG*
- Call Server 18
- Call Session Control Function *See CSCF*
- Campus network *See NPN*
- CAP 10
- C-band 131
- CBRS 130
- Cell
 - free 299
 - less 299
- Centralized-RAN *See C-RAN*
- Channel coding 148
- CHF 181
- Circuit Switched Fallback *See CSFB*
- Circuit Switching 7
- Cloud 262
 - Central 226
 - Computing 69
 - Edge 226, 242
 - Security 236
- Cloudification 134
- Cloud-RAN *See C-RAN*
- Cobot 288
- Codewords 148
- Compute as a Service *See CaaS*
- Conference Server 18
- Connection 5
- Connectionless communication 6
- Connection-oriented communication 4
- Control Plane 75, 137, 159, 162, 167, 238
- Control Unit 159
- CPE 213
- C-RAN 72, 160, 226
- CRC 147
- CRG 215
- CSCF 33
 - Emergency 36
 - Interrogating 33
 - Proxy 33
 - Serving 33

- CSFB 61
- Customer Premises Equipment *See CPE*
- Data center 74, 226, 262
- Data Plane 75, 238
- Data-Controller Plane Interface *See D-CPI*
- Datagram 7
- D-band 283
- D-CPI 83
- Deep packet inspection 236
- Denial of Service *See DoS*
- Determinism 277
- Diameter 46
 - Application 46
 - Base Protocol 46
 - Messages 48, 50, 52
 - SIP Application 49
- Diameter Signaling Controller *See DSC*
- Digital twin 271, 273, 287
 - Network 300
- Distributed Unit 159
- DoS 232, 234, 238, 241
- Dose 256
- Drones 105, 107, 217, 302
- DSC 181
- DSL 213
- Dual Connectivity 159
- EAP 216, 248
 - Authenticator 248
 - Client 248
 - Server 248
- EBS 130
- eHealth 288
- EIR 11, 181
- eMBB 103, 105, 109, 113
- Emergency call 15, 31, 36
- Emission 255
- eNB 59, 158
- Encryption 235
- Energy
 - Consumption 261f.
 - Efficiency 260f.
 - Harvesting 298
 - Renewable 262
 - Requirements 259
- en-gNB 158
- Enhanced Mobile Broadband *See eMBB*
- EPC 60, 158, 204
 - E-UTRAN 59, 204
 - eV2X 105
 - Evidence 251
 - Evolved Packet Core *See EPC*
 - Exposure 249, 252, 255f., 258
 - Extended Reality *See XR*
 - FCAPS 165
 - FCC 130
 - FDD *See Frequency Division Duplex*
 - Features 135, 268
 - Femtocell 250, 258, 260
 - FFT 151
 - FG NET-2030 271
 - Firewall 235
 - Fixed Mobile Convergence *See FMC*
 - Fixed wireless access 213
 - Fixed-Mobile Interworking Function *See FMIF*
 - Flow table 77f., 83
 - FMC 211, 213, 226
 - Scenarios 214
 - FMIF 214
 - Frame 153
 - Frame structure 152
 - Frequency auction
 - Austria 129
 - Germany 128
 - Switzerland 128
 - USA 130
 - Frequency Division Duplex 146
 - Frequency ranges 121
 - EU 129
 - FR1 121f.
 - FR2 121f.
 - USA 130
 - Worldwide 131
 - Fronthaul 72, 227
 - Interface 164, 166
 - Future Network 95, 204f.
 - FWA *See Fixed wireless access*
 - Gateway 18, 63
 - Access 15
 - Decomposed 18, 40
 - Media 18
 - Residential 15
 - Signalling 18
 - Trunking 16
 - Generic Routing Encapsulation *See GRE*

- GERAN 11, 61, 204
gNB 158, 169, 218, 247
GPRS 12
– Core 204
gPTP 222
GRE 75
GSM 10
– Core 204
GTP-U 162, 174, 176
- H.248 40
Handover 63, 207, 247
HAPS 217
Haptic 271
HARQ 146
Health 251, 254, 257, 263
Hexa-X 280f., 289, 301
High Altitude Platform Station *See* HAPS
High Layer Split 161
HLR 11, 32
Hologram 271
Holographic Radio *See* HR
Holographic-Type Communications *See* HTC
Holography 271, 275
Home Environment 243
Home Subscriber Server *See* HSS
HR 299
HSS 32, 207, 210
HTC 272f., 286
HTTP/2 193, 247
Hypervisor 66, 227
- IaaS 69, 196
IARC 251
IBCF 36, 210
ICNIRP 250
IEEE 251
IFFT 151
Immission 255, 257
– Limits 257
IMS 32, 204, 209
IMT-2020 103, 139
IMT-2030 *See* 6G
– Promotion Group 282
– Requirements 293ff.
– Usage scenarios 290
INAP 10
Industry 4.0 107
Infrastructure as a Service *See* IaaS
- Interconnection Border Control Function *See* IBCF
International Agency for Research on Cancer
See IARC
International Commission on Non-Ionizing Radiation Protection *See* ICNIRP
Internet of Things *See* IoT
Intersymbol interference 155
Interworking
– 4G/5G 207
Intrusion
– Detection 235
– Prevention 235
IoT 219
– Broadband 220
– Cellular 220
– Critical 221
– Definition 219
– Industrial automation 222
– Massive 220
IP Multimedia Subsystem *See* IMS
IPsec 247
ISUP 10
- JavaScript Object Notation *See* JSON
JSON 185, 192
- Key 246
– Session 247
– Shared 246
– Temporary 247
Key Performance Indicator *See* KPI
Key Value Indicator *See* KVI
KPI 114, 275, 293f.
KV enabler 293
KVI 293
- Latency 99, 102, 106, 112, 272, 275
Lawful Interception 15, 60
Layer 148
LDPC 148
Lifecycle 198, 202, 229, 237
LINP 97
Location Server 18
Logically Isolated Network Partition *See* LINP
Low Layer Split 161
LTE 30, 158, 204
LTE for Machines *See* LTE-M
LTE-M 220, 222

- Macrocell 250, 258, 261
- Malicious Apps 232
- Malware 232, 234
- Man-in-the-middle 232, 244
- MANO 226, 228
- ManyNets 273
- MAP 10
- Massive Machine Type Communications *See mMTC*
- Master Node 159
- MEC 73, 226, 242
 - Reference architectur 74
- Media
 - Gateway 13
 - Gateway Controller 18
 - Resource Function 36
- Media Gateway Control Function *See MGCF*
- Megaco *See H.248*
- Metadata 93, 278
- METIS 99
- MGCF 34
- Microcell 250, 258, 260
- Migration 206
 - 4G/5G 206
- Millimeter waves *See mm waves*
- MIMO 149, 154, 249, 260
 - Extreme 299
 - Multi User 155
 - Single User 155
- Mixed Reality *See MR*
- mm waves 122, 125, 130, 213, 249, 253
- MME 60, 207
- mMTC 103, 105, 109, 113, 230
- Mobile networks 2, 10, 14, 30, 204, 250
- Mobile radio radiation 251
- Mobility Management Entity *See MME*
- Modularization 133
- Modulation 154
- MPLS 9, 75
- MPTCP 217, 260
- MR 286
- MSC 10
- MTP 18
- Multi Sensing 287
- Multi-access Edge Computing *See MEC*
- Multimedia over IP 209
- Multiple Input Multiple Output *See MIMO*
- Multiprotocol Label Switching *See MPLS*
- N3IWF 212
- N5CW 216
- NaaS 69, 196
- Narrowband IoT *See NB-IoT*
- NAS 173, 200
 - MM 173
 - SM 173
- NBI 239
- NB-IoT 220, 222
- NEF 181, 194, 248
- Network 2030 271, 276
 - Architectural principles 276f.
 - Capabilities 275
 - Functional areas 278
 - Performance 275
 - Requirements 273, 275
 - Use cases 273
 - Verticals 274
 - Vision 274
- Network as a Service *See NaaS*
- Network functions 138
- Network Functions Virtualisation *See NFV*
- Network Service Header 93
- Network Slice Selection Function *See NSSF*
- Network slicing 138, 143, 195, 198, 200, 228
 - Generic Slice Template 199
 - Network Slice Type 199
 - NSI 199
 - NSSAI 199, 201
 - NSSI AN 202
 - NSSI CN 202
 - Slice Differentiator 199
 - Slice/Service Type 199
 - S-NSSAI 199f., 202
- Next G Alliance 281, 297
- Next Generation Network *See NGN*
- NFV 65f., 72, 228
 - Framework 69
 - Management and Orchestration 68
 - MANO 68
 - NFVI 67
- NFVI as a Service *See NFVaaS*
- NFVaaS 70
- NFVO 228
- NG protocol stacks 162
- NGAP 162, 172
- ng-eNB 158
- NGMN 142
 - Alliance 101

- NGN 15
- NIR 255, 259
- Non Access Stratum *See NAS*
- Non-3GPP access 211, 259
- Non-5G-Capable over WLAN *See N5CW*
- Non-ionizing radiation *See NIR*
- Non-Public Network *See NPN*
- Non-standalone system 136
- Non-Terrestrial Network Gateway *See NTN GW*
- NPN 223
 - Hybrid 224
 - Standalone 223
 - Variants 224
- NR 122, 139, 145, 149, 158, 260
- NRF 181, 183, 201, 247
- NSA *See Non-standalone system*
- NSA architecture 136
- NSSAAF 180
- NSSF 180, 201f.
- NTN GW 218
- NWDAF 181
- OAM 299
- OAuth procedure 247
- O-Cloud 165
- OFDM 149f.
 - CP 152
 - F 152
 - Symbol 151f.
- OFDMA 150
- one6G 282
- ONF 89
- ONIR 251
- Open Networking Foundation *See ONF*
- Open RAN *See O-RAN*
- Open source software 163
- OpenFlow 77
 - Alternatives 89
 - Channel 77
 - Messages 83, 86
 - Packet processing 82
 - Session 85
 - Switch 78
- Open-RAN *See O-RAN*
- O-RAN 163
 - Alliance 163
 - Reference architecture 163
 - Use cases 165
- Orbital Angular Momentum *See OAM*
- Orchestration 65, 72f., 201
- Ordinance on Protection from Non-ionising Radiation *See ONIR*
- OSS 228
- OTT 61
- Over The Top *See OTT*
- Packet Data Network-GW *See PDN-GW*
- Packet filter 235
- Packet Switching 8
 - Datagram 9
 - Virtual Circuit 8
- Paging 162
- PAL 130
- Passive Optical Network *See PON*
- PBCH 146
- PCF 180, 207
- PCRF 60, 207
- PDCCH 146
- PDCP 171
- PDN connection 207
- PDN-GW 60
- PDSCH 146f., 152
- PDU session 184, 195, 200, 207, 210, 217
- PFCP 174
 - Session contexts 174
- PGW 207
- Physical layer 145
- Physical Network Function *See PNF*
- Picocell 250, 258, 260
- Pishing 241
- PKI 242
- Plant limits 257
- Platooning 110
- PNF 197, 227
- Policy and Charging Rules Function *See PCRF*
- PON 213, 226
- PRACH 146
- Precautionary principle 257
- Priority Access License *See PAL*
- Programmability 276
- Protocol stacks
 - Control plane 170
 - User plane 176, 179
- Proxy Server 17
- Public Key Infrastructure *See PKI*
- PUCCH 146
- PUSCH 146

- QAM 148f.
- QPSK 148f.
 - Modulation 149
- Quantum
 - Computing 278, 298
- Radio Access Network *See RAN*
- Radio cell 250
- Radio channel 154
- Radio Transmission 145
- Radio Unit 159
- RAN 59, 72, 145, 157, 205, 217, 259
 - Architecture 136, 159, 161
 - Migration 208
 - NG 137
 - Options 157
 - Satellite 217
 - Sharing 163
 - Slicing 201
 - Splitting 161
- RAN Intelligent Controller *See RIC*
- Ransome Ware 232
- rAPP 165
- RAT 157
- Raw materials 262
- Reconfigurable Intelligent Surfaces *See RIS*
- RedCap 140, 269
- Registrar Server 17
- Regulation 127
 - Bundesnetzagentur 127f.
 - Eidgenössische Kommunikationskommission 127
 - FCC 130
 - Telekom-Control-Kommission 127
- Remote Radio Head *See RRH*
- Requirements 99, 101, 267
 - 3GPP 105, 115
 - ITU-R 103, 112
 - METIS 100
 - NGMN 101
- Residential Gateway 213
- Resource
 - Block 149, 154
 - Element 153
- RESTful 192
 - API 193
- RG *See Residential Gateway*
- RIC 163
 - Near-RT 165
 - Non-RT 165
- RIS 299
- RLC 171
- Roaming 40, 190, 246
- Router 78, 226
- RRC 171
- RRH 72
- RTP 28, 63
 - Packet 28
 - Session 28
- RU 218, 259
- SA *See Standalone*
- SA architecture 136
- SAR 256
- SAS 130
- Satellites 217
- SBA 137, 182f., 228
- SBI 183, 192, 239
 - Protocoll stack 193
- SCP 181
- Scrambling 148
- SCTP 18, 172
- SDAP 176
- SDG 263, 292
 - Mobile networks 263
- SDN 75, 97
 - Application 77
 - Architecture 76f.
 - Controller 75, 77, 227
 - Switch 75, 77, 226
 - Use cases 91
- SDP 27
 - Message 22
 - Offer/Answer model 27
 - Parameter 27
- Secondary Node 159
- Security 230, 233
 - Application 239
 - Application Domain 243
 - Architecture 242
 - Areas 233
 - Authentication 242
 - Automation 236
 - Cloud 230, 236, 241
 - Encryption 242
 - Framework 233
 - Hypervisor 236
 - Monitoring 242

- Multi-factor authentication 242
- Network Access 242
- Network Domain 243
- Network operation 234
- Network slices 240
- NFV 236
- O-RAN 235
- Requirements 234, 243
- Requirements for 5G 244
- SBA 247
- SBA Domain 243
- SDN 230, 238
- Stakeholder 231
- System components 230
- Technologies 238
- Threats 232
- User Domain 243
- Virtualization 230
- Visibility and Configurability 243
- Zone 235
- Security Edge Protection Proxy *See SEPP*
- Self Interference Cancellation *See SIC*
- Sensor technology 287f.
- SEPP 181, 191, 247
- Service
 - Discovery 194
 - Registration 194
- Service Based Architecture *See SBA*
- Service Based Interface *See SBI*
- Service Chaining 90
- Service Function 91
 - Chain 197
 - Chaining 91
 - Forwarder 92
 - Path 92
- Serving network 242
- Serving-GW *See S-GW*
- Session Border Controller 18, 235
- SFC 91
 - Controller 92
- SGW 207
- S-GW 60, 207
- SIC 299
- Sidelink communication 297
- Signaling Transfer Point *See STP*
- SIGTRAN 13
- Single Radio Voice Call Continuity *See SRVCC*
- SIP 19, 37
 - Header 23
- Messages 20, 22
- Registrar Server 20
- Registration 37, 51
- Request 20
- Response 20
- Routing 24
- Session 21, 39
- Three-Way Handshake 21
- Transaction 23
- SIP URI
 - Permanent 20
 - Temporary 20
- SKV 293
- Sleep mode 260
- Small cell 260
- SMF 170, 184, 207, 248
- SMO 165
- SMSF 173, 182
- SNS JU 281
- Social Key Value *See SKV*
- Socialized Internet of Things 273
- Software Defined Networking *See SDN*
- Softwarization 133, 195
- Space communication 273
- Special Purpose Network *See SPN*
- Spectrum 283
 - Clearing 284
 - Efficiency 293
 - Exploration 284
 - High-band 283
 - Low-band 283
 - Mid-band 283
 - Refarming 284
 - Sharing 298
 - THz 283
- Split computing 305
- SPN 304
- Spoofing
 - DNS 233
 - IP 233f.
- Spyware 232
- SRI 218
- SRVCC 63
- Stakeholder 231
- Standalone RAN 136
- Standardization 125
 - 3GPP 125
 - ITU-R 125
 - ITU-T 125

- STP 181
- Stream 148, 155
- Subcarrier 151, 154
- Sustainability 262
 - 5G 265
- Sustainable Development Goal *See SDG*
- Synchronization 272, 275
 - Tactile 271, 273, 275
- TDD *See Time Division Duplex*
- TDoS 234
- Telecommunications Standards Development Society, India *See TSDSI*
- Telephone Denial of Service *See TDoS*
- Tenants 133, 167, 195, 237, 241
- Time Division Duplex 146
- Time division multiplex
 - Asynchronous 6
 - Synchronous 8
- Time Sensitive Networking *See TSN*
- TNAN 212
- TNAP 212
- TNGF 212
- Transaction 23, 41
- Transport network 226
- TrGW 210
- Trusted Non-3GPP access 212
- Trusted WLAN Access Point *See TWAP*
- Trusted WLAN Interworking Function *See TWIF*
- TSDSI 282
- TSN 222
 - Bridge 222
 - Translator 222
- TUP 10
- TWAP 216
- TWIF 216
- UAS 217
- UCMF 181
- UDM 180, 207, 210, 246
- UDR 180f., 191
- UDSF 181, 191
- Ultra-Reliable and Low Latency Communications
 - See* URLLC
- UMTS 12, 204
- Unified Data Repository *See UDR*
- Uniform Resource Identifier *See URI*
- Unmanned Aerial System *See UAS*
- Unstructured Data Storage Function *See UDSF*
- Untrusted Non-3GPP access 211
- UPF 170, 179, 207
 - I 179
 - PSA 179
- URI 20, 192
- URLLC 103, 105, 109, 113, 230
- Use case 99, 111
 - 3GPP 105
 - 5GPPP 104
 - V2X 109
- User Agent 17
- User Plane 75, 137, 159, 162, 167
- USIM 242
- UTRAN 59, 61, 204
- UVEK 251
- Verticals 105, 231, 274
- VIM 69, 229
- Virtual Extensible LAN *See VXLAN*
- Virtual Network Function *See VNF*
- Virtual networks 143
- Virtual Reality *See VR*
- Virtualization 66, 72, 97, 160, 227
- Virtualized Infrastructure Manager *See VIM*
- VLAN 75
- VLR 11
- VNF 196, 227
 - Forwarding Graph 71, 197
 - Instance 68
 - Set 68
- VNF Manager *See VNFM*
- VNFM 68, 228
- Voice over IP *See VoIP*
- Voice over LTE *See VoLTE*
- Voice over Wifi *See VoWiFi*
- VoIP 19, 63, 209
- VoLGA 62
- VoLTE 61f.
- VoWiFi 64, 211
- VR 286
- VXLAN 75
- W-5GAN 215
- W-5GBAN 215
- W-5GCAN 215
- W-AGF 215
- W-band 283
- WHO 255

- Wireline 214
 - 5G Access Network 215
 - 5G Cable Access Network 215
 - Access Gateway Function 215
- World Health Organization *See WHO*
- World Radiocommunication Conference 121,
282
- WRC *See World Radiocommunication
Conference*
- xApp 165
- Xn protocol stacks 162
- XnAP 163
- XR 268, 286