Ulrich Trick

5G

Also of interest

The 5G Myth
When Vision Decoupled from Reality
William Webb, 2019
ISBN 978-1-5474-1728-5, e-ISBN (PDF) 978-1-5474-0118-5,
e-ISBN (EPUB) 978-1-5474-0120-8

Digital Electronic Circuits
Principles and Practices
Shuqin Lou, Chunling Yang, 2019
ISBN 978-3-11-061466-4, e-ISBN (PDF) 978-3-11-061491-6,
e-ISBN (EPUB) 978-3-11-061493-0

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
e-ISBN (EPUB) 978-3-11-066669-4

Communication, Signal Processing & Information Technology
Series: Advances in Systems, Signals and Devices, 12
Edited by Faouzi Derbel, Nabil Derbel, Olfa Kanoun, 2020
ISBN 978-3-11-059120-0, e-ISBN (PDF) 978-3-11-059400-3,
e-ISBN (EPUB) 978-3-11-059206-1

Signal Processing and Data Analysis
Tianshuang Qiu, Ying Guo, 2018
ISBN 978-3-11-046158-9, e-ISBN (PDF) 978-3-11-046508-2,
e-ISBN (EPUB) 978-3-11-046513-6

Ulrich Trick

5G

An Introduction to the 5th Generation Mobile Networks

DE GRUYTER
OLDENBOURG

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

ISBN 978-3-11-072437-0
e-ISBN (PDF) 978-3-11-072450-9
e-ISBN (EPUB) 978-3-11-072462-2

Library of Congress Control Number: 2020951340

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.

Cover illustration: PocholoCalapre/iStock/Getty Images Plus, Ani_Ka/DigitalVision Vectors/Getty Images
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

To Cornelia

To Susanne, Johanna, and Thomas

To Mafalda, and the memory of Werner

Preface

With 5G, mobile networks and telecommunications networks have entered a new phase. 5G mobile networks use unique concepts and technologies to deliver current and future applications across a wired spectrum, from high bit-rate smartphones to high-availability car-to-x and mass IoT applications.

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) incl. 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 approach to 5G is different from previous versions, which were mainly driven by technology. Chapter 4 shows that at the beginning of 5G, there were 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. 5G systems are currently being introduced according to Release 15; Release 16 is currently in progress, while Release 17 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, it has identified possible frequency ranges for 5G. These, in turn, have been and are still allocated to the network operators by regulators.

Chapter 6 provides an overview of a 5G system based on the applied design principles, the implementation features, and associated functions, and the resulting network architecture.

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).

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.

https://doi.org/10.1515/9783110724509-202

Chapter 9 summarizes the previously introduced concepts in an overall view, taking into account 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 it also represents a new generation converged network.

Since this is still an IP network, we have to 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.

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. Chapter 12 finally gives an outlook on the future with the further development of 5G at 3GPP, the work on a network 2030 at the ITU, and first considerations on 6G in different research projects.

Compared to the German-language edition, more extensive extensions were made in Section 5.3 on international regulation and Section 12.2 on the progress made in the specification of a Network 2030. Besides, necessary updates and additions were made to a small extent.

The book's main objective is to provide people interested in 5G technology and application scenarios with a well-founded knowledge for an introduction to 5G and encourage further discussion of this topic. For this, the book refers to numerous additional sources. This group of people includes persons with a general interest in technology, mostly 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.

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

I would especially like to thank Dr. Gerd Zimmermann, a proven 5G, and experienced 3GPP standardization expert. He supported me with his expertise and references to sources and advised me on the book's content and structure. Significantly, Chapters 4 to 9 have benefited from his friendly support. Interestingly, the continuing collaboration on 5G with Gerd Zimmermann has completed a circle that began in 1987 at the Department of Communications Engineering at the University of Kaiserslautern. He was a prospective, and I was an advanced doctoral student under Prof. Dr. Werner Rupprecht. One of the seeds laid at that time has come to fruition with this 5G book.

I would also like to take this opportunity to thank another long-time companion, Prof. Dr. Armin Lehmann, for his contribution to this book with his valuable

expert input, especially in Chapters 2, 3, and 10, in addition, for his friendship and technical support in all aspects of telecommunications networks.

I want to thank the already mentioned Gerd Zimmermann and Armin Lehmann for their critical review of the manuscript, together with many other suggestions and essential tips.

With regards to 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 never have been published. In fact, this book project was the very beginning of this fruitful collaboration between Seattle and Frankfurt. Thanks a lot! I would also like to thank M.Sc. Susanne Trick for her numerous additional notes on the linguistic design of this book.

Last but not least, I would like to thank the De Gruyter publishing house. Mr. Leonardo Milla suggested the book topic 5G to me; Mrs. Ute Skambraks was again very supportive in the creation; Mrs. Elisabeth Stanciu accompanied the work on this book up to the production. Many thanks to all of you for your support, excellent cooperation, and the extremely well-designed cover.

Contents

1 Evolution of Mobile Networks

With 5G, the development of mobile networks has entered a new phase. So far, 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 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 [187].

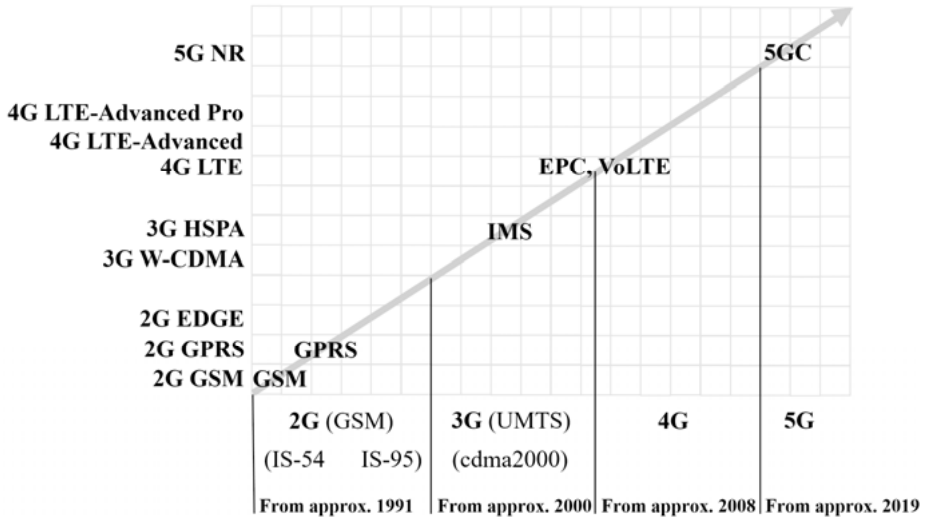

Fig. 1.1: Evolution of mobile networks [54]

https://doi.org/10.1515/9783110724509-001

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) [186] standardized the 3G cdma2000 solution with several successive versions [187].

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.

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. 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 and energy consumption. Finally, we take a look into the future, first at the further development of 5G and then at an already planned 6th generation. 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.

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 demonstrated in full. 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 the characterization of telecommunications networks, are worked out below.

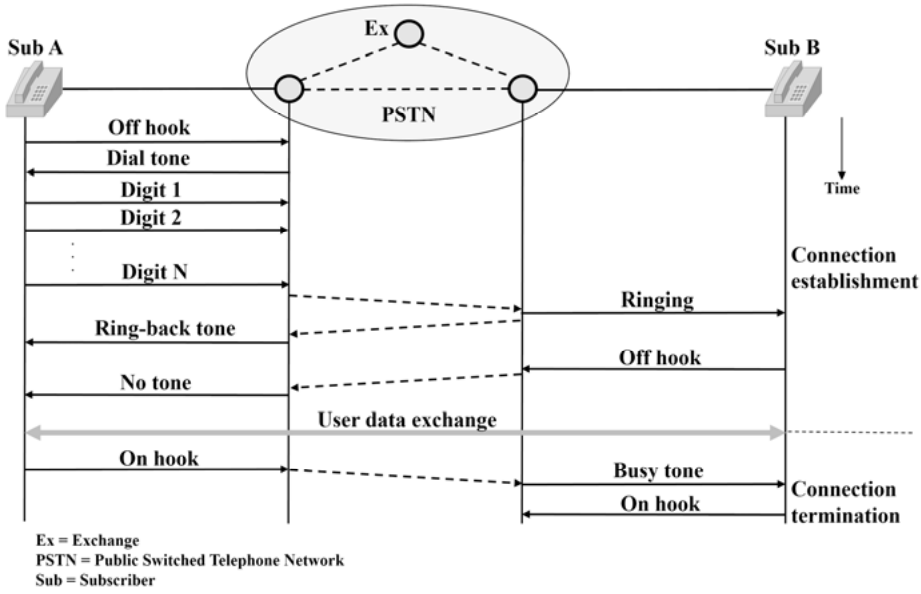

Fig. 1.2: Connection-oriented communication using the example of a telephone call in a PSTN

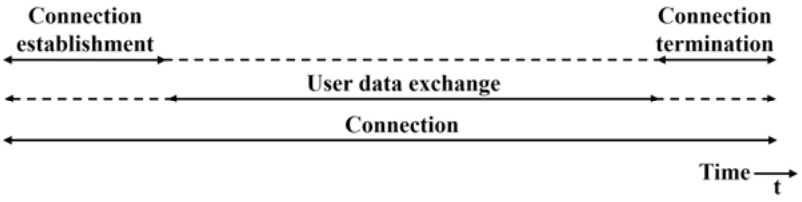

Fig. 1.3: The term connection in a generally applicable representation

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.

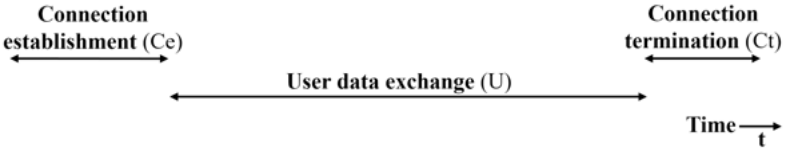

Fig. 1.4: Connection-oriented communication with a physically switched circuit

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.

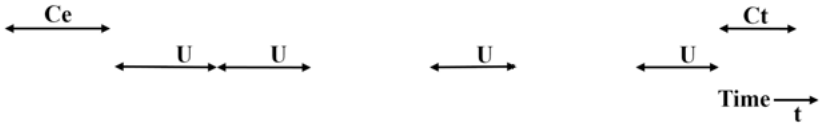

Fig. 1.5: Connection-oriented communication with virtual channel

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 for 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.

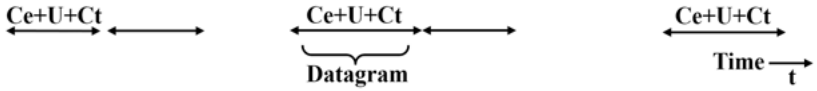

Fig. 1.7: Connectionless communication

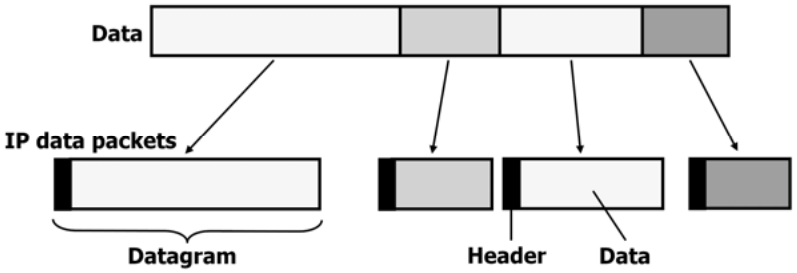

Fig. 1.8: Generation of datagrams or IP packets

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 application 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.

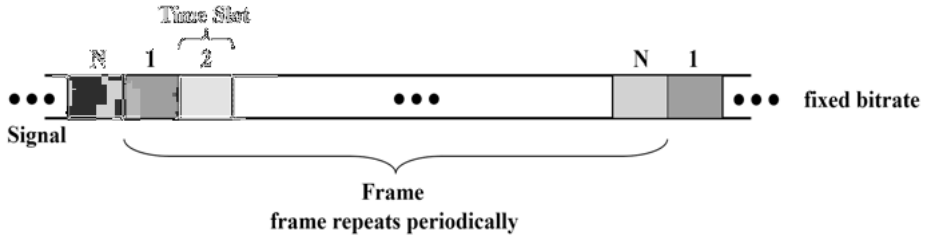

Fig. 1.9: Synchronous time division multiplex

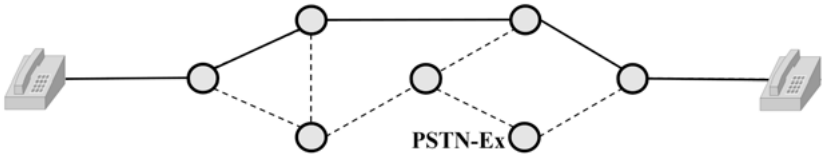

Fig. 1.10: Routing principle Circuit Switching using the example of PSTN

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, the 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].

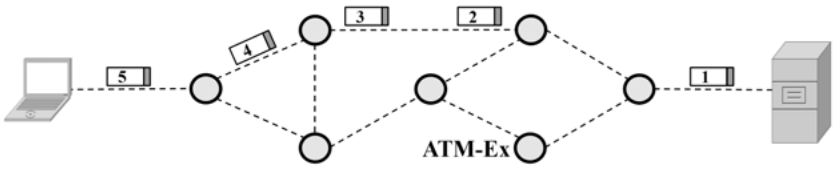

Fig. 1.11: Routing principle Virtual Circuit Packet Switching using the example of an ATM network

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].

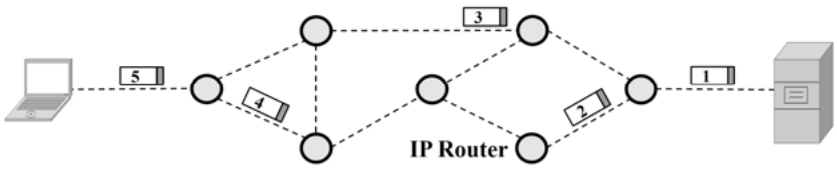

Fig. 1.12: Routing principle Datagram Packet Switching using the example of an IP network

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 a 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).

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 BTSs.

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].

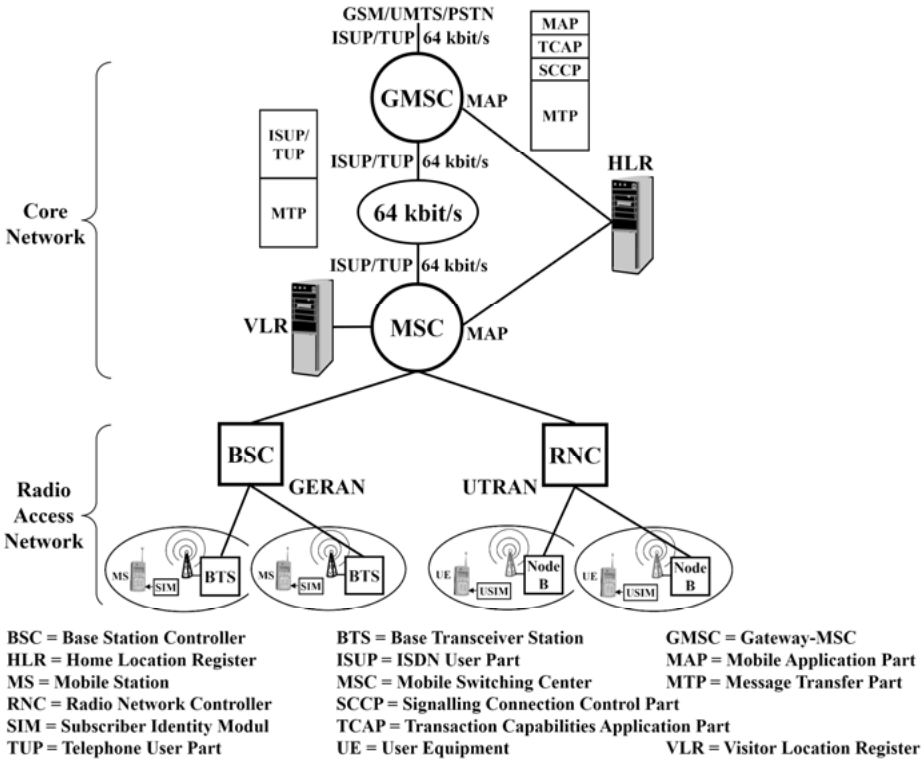

Fig. 1.13: 3GPP Release 99 UMTS mobile network with the focus on the circuit-switched GSM core network

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 switch-

ing 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].

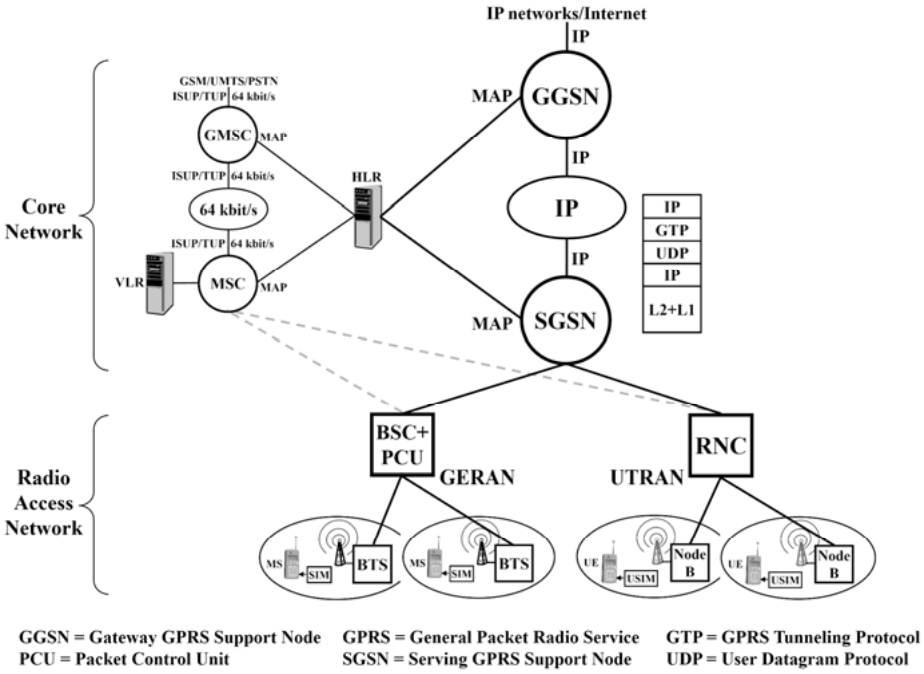

Fig. 1.14: 3GPP Release 99 UMTS mobile network with GPRS core network

As already mentioned above, the next evolutionary step with UMTS 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].

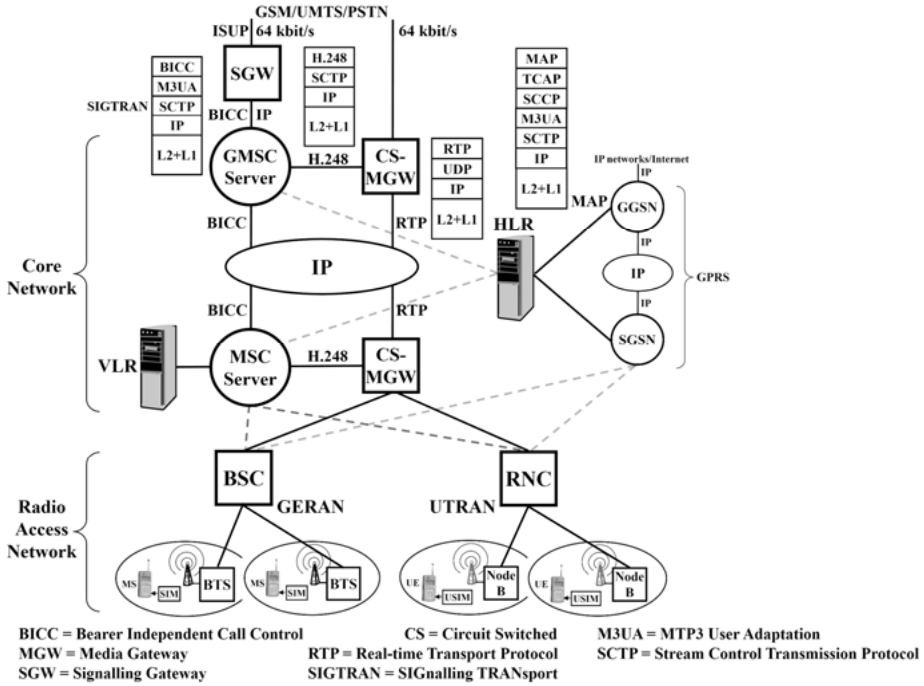

Fig. 1.15: 3GPP Release 4 UMTS mobile network

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. UMTS Release 99: Circuit Switching for voice and Datagram Packet Switching for data
4. UMTS 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 the 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 connects a circuit-switched (e.g., ISDN) and a packet-switched (PS) core network (trunking gateway).

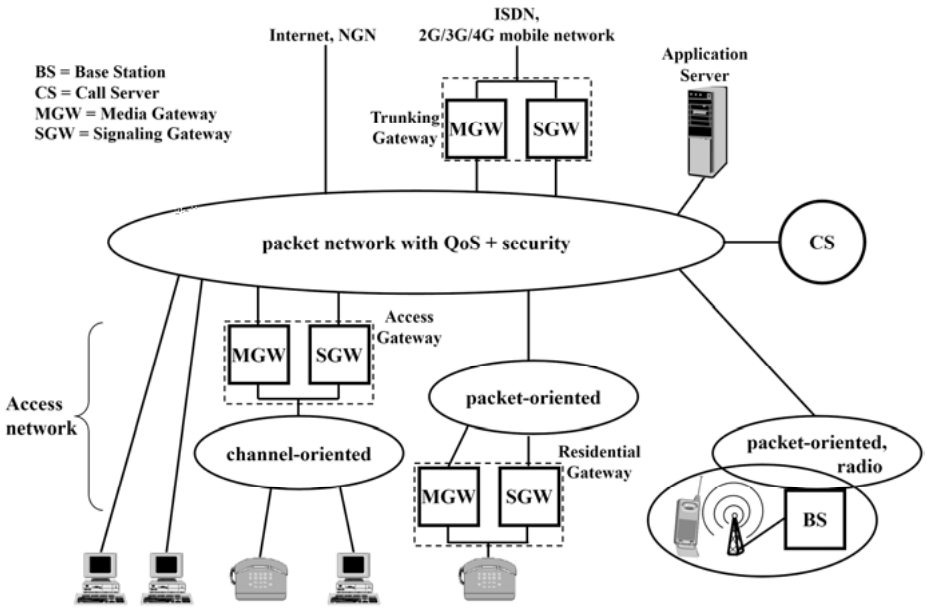

Fig. 1.16: Basic network structure of an NGN

The ITU-T (International Telecommunication Union-Telecommunication Standardization Sector) describes the concept “Next Generation Network (NGN)” in its defini-

tion 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 the existing 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.

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.

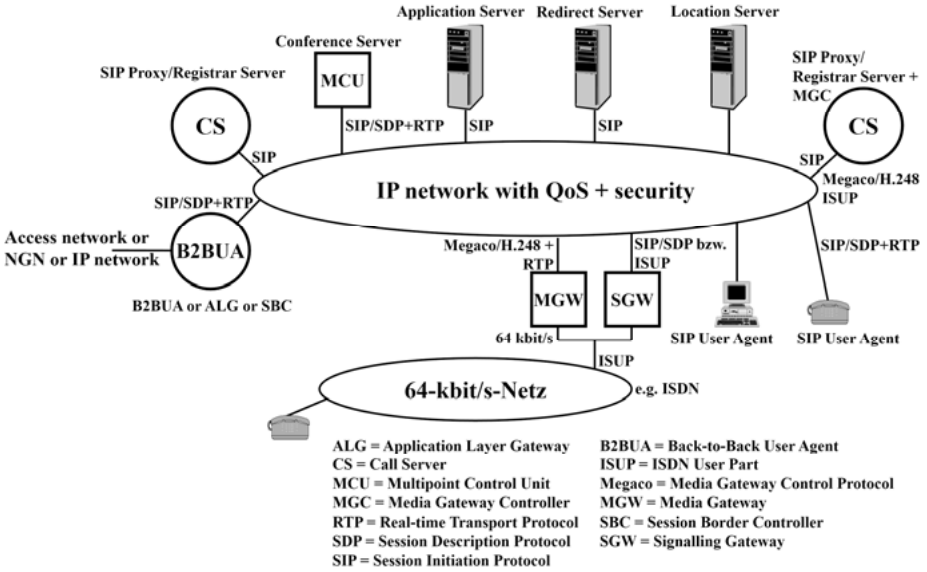

Fig. 1.17: Protocols and network architecture for NGN with SIP for signaling

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 the 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) are representing 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.

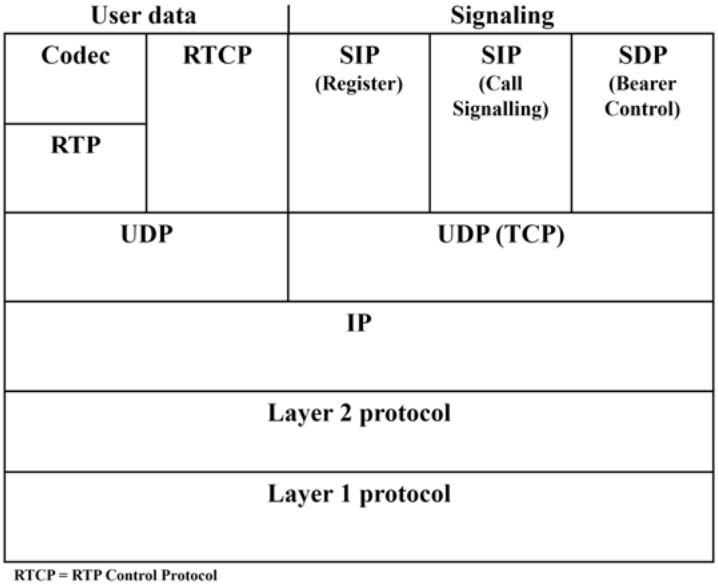

Fig. 1.18: Protocol stacks for Multimedia over IP

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, which is crucial for SIP routing, is determined in the SIP registration process shown 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 request 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 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.

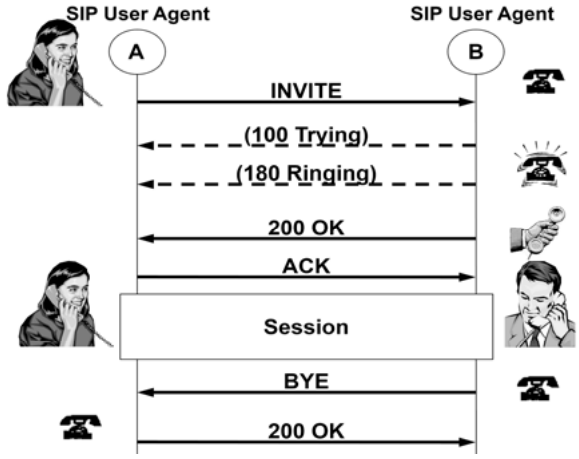

Fig. 1.19: Typical message exchange during SIP session setup and termination

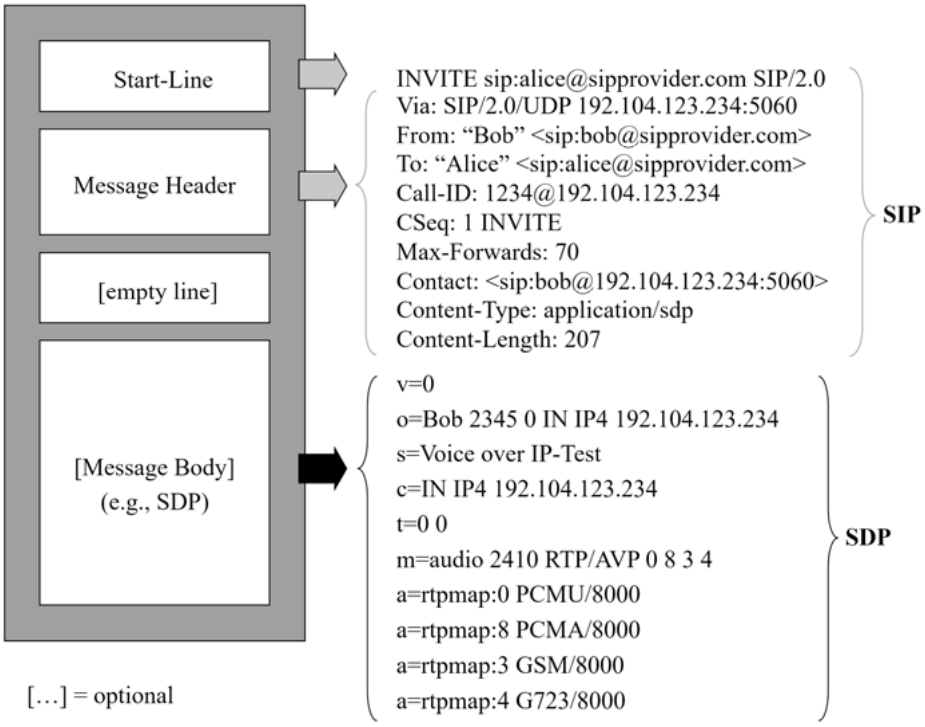

Fig. 1.20: Structure of SIP messages with message header and message body

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
Via	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
From	Permanent SIP-URI of the request sender, the transaction initiator, the so-called User Agent Client (UAC)
To	Permanent SIP-URI of the request receiver. Replies with response(s), so-called User Agent Server (UAS)
Call-ID	Identifies all SIP messages that belong to a session or dialog
CSeq	Identifies all SIP messages that belong to a transaction (request + all resulting responses)
Max-Forwards	Number of still allowed SIP hops to avoid endless loops
Contact	Temporary SIP-URI of the SIP User Agent sending the message
Content-Type	Data type in the message body, here SDP
Content-Length	Data length in Byte in the message body
Expires	Validity period in seconds of a SIP event, such as a registration
Allow	Methods supported by the SIP network element
User-Agent	Description of the User Agent
Record-Route	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
Route	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

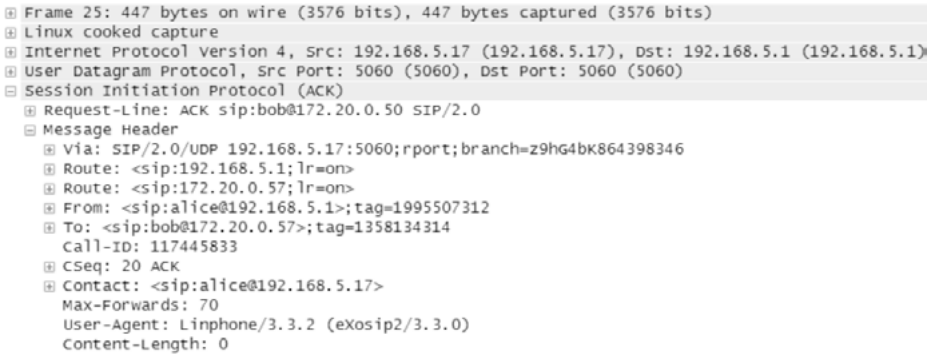

Fig. 1.22: SIP ACK request captured with protocol analysis software

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 subscriber 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.

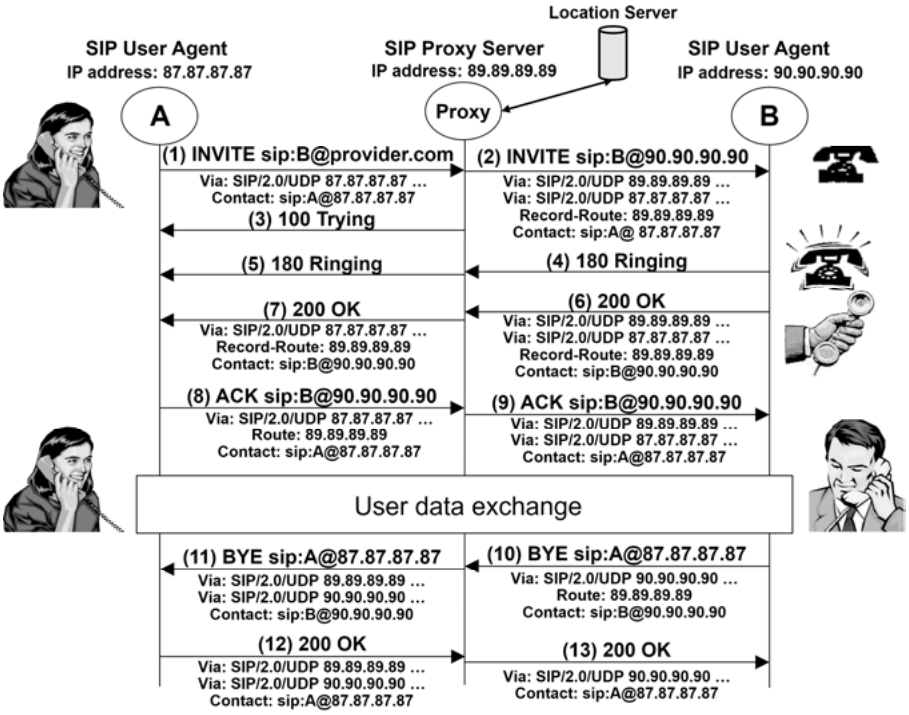

Fig. 1.23: Routing of SIP messages

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 his 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=reconly (receive only) a=sendrec (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: 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, Figure 1.26, a corresponding protocol record from network practice [173; 5].

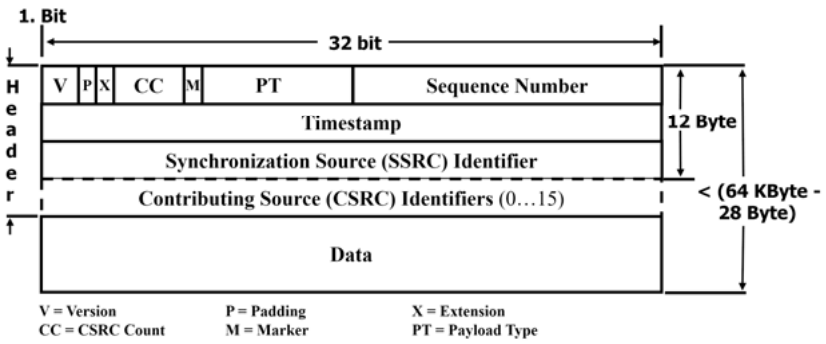

Fig. 1.25: Structure of an RTP packet

```
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: dabb7fbed85cde7abaffbff1ffdf5efad77edefeb5f4b2f6...
```

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 the 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 offers bit rates of up to 100 Mbit/s per radio cell.

The third and so far last revolutionary step in 3GPP mobile networks will be 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 17, we will then 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**
– 2000	– 2001
– Core network same as GSM + GPRS	– Separation of signaling and user data in the core network
– Access network UTRAN + GERAN	– Instead of MSC MSC-Server + MGWs
– Higher data rates, up to 2 Mbit/s	– CCS no.7 (Common Channel Signaling) over SIGTRAN
– USIM (UMTS Subscriber Identity Module)	– QoS architecture for PS Domain
– AMR Codec (Adaptive Multi-Rate), 3,4 kHz	

https://doi.org/10.1515/9783110724509-002

Release 5
- 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

Release 6
- 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
- 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

Release 8
- 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
- 2009
- Self-Organizing Networks (SON).

Release 10
- 2011
- Network optimization for M2M
- Carrier Aggregation
- LTE Advanced, up to 1000/500 Mbit/s down-/ upstream

Release 11
- 2012
- EVS Codec (Enhanced Voice Services)
- WebRTC (Web Real-Time Communication between Browsers) with IMS
- Proximity-based Services (ProSe) with Device-to-Device communication

Release 12
- 2015
- LTE-M (LTE for Machines) for M2M and IoT (Internet of Things)

Release 13
- 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

Release 14
- 2017
- V2X (Vehicle to X)
- Virtualization, Orchestration

Release 15
- June 2019
- 5G Phase 1

Release 16
- Planned December 2020
- 5G Phase 2

Release 17
- Planned March 2022
- 5G further development

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.

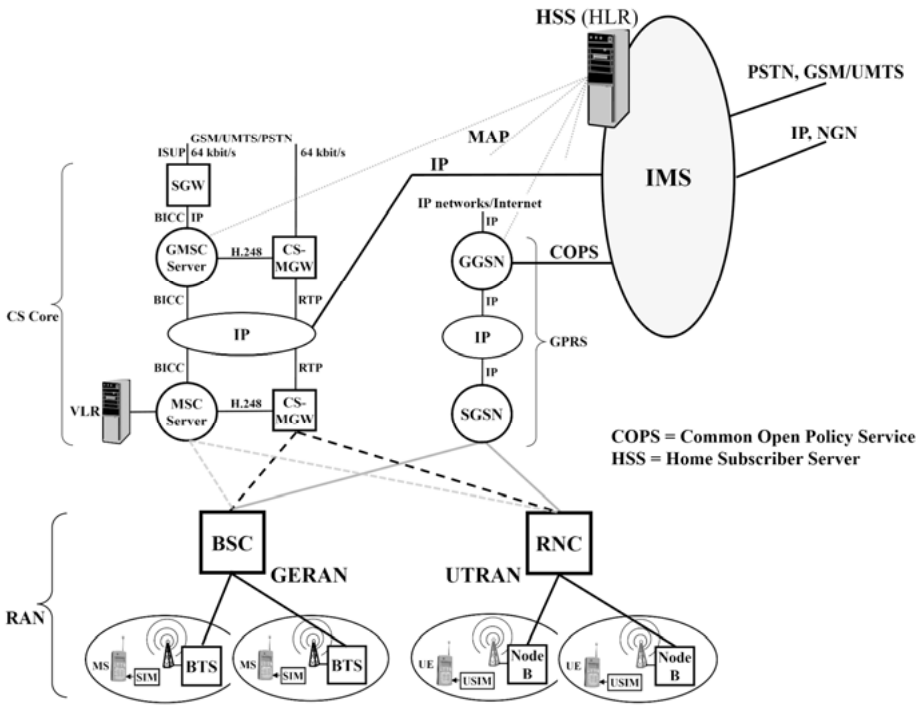

Fig. 2.1: 3GPP Release 5 mobile network

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.

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 will be discussed in more detail below. Concerning timeliness, the statements are based on 3GPP Release 8.

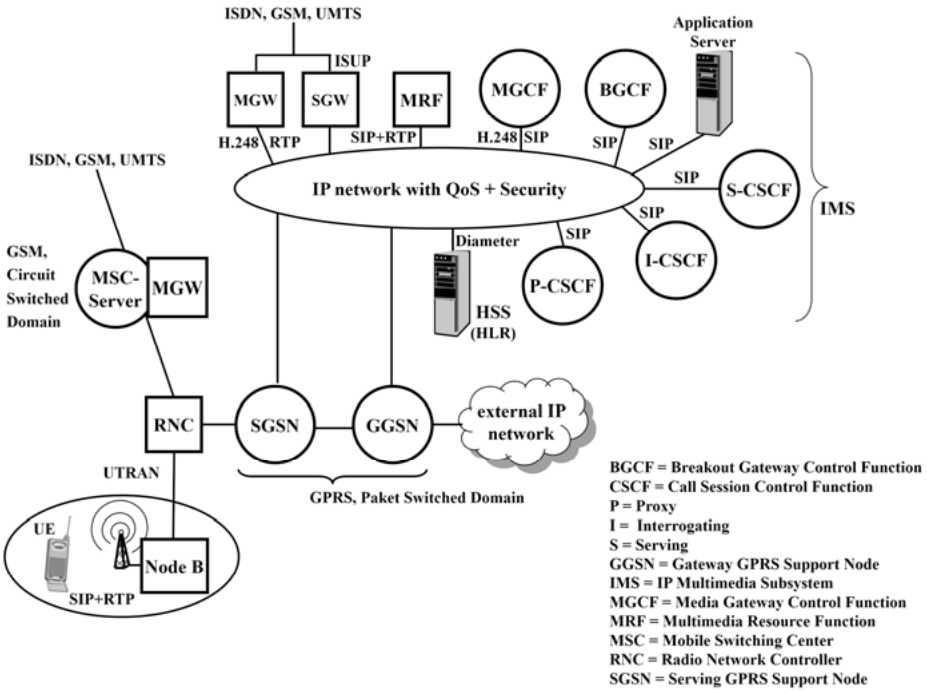

Fig. 2.2: IMS in a 3GPP Release 5 mobile network

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).

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 IMSs 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) 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 (Hypertext 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, transcoding of multimedia data.

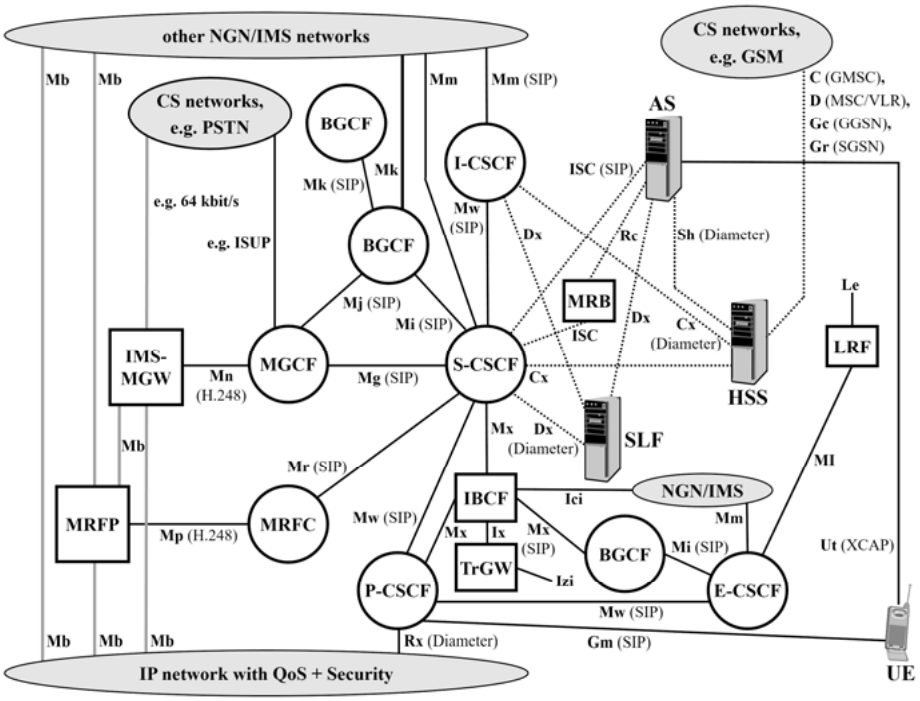

Fig. 2.3: Network elements, reference points, and protocols in IMS in 3GPP Release 8 [30]

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 necessary 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 with the focus on the registration process in IMS. 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.

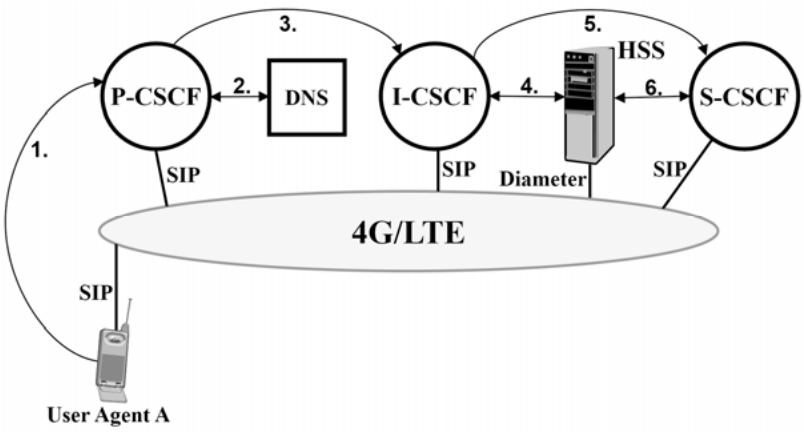

Fig. 2.4: Network architecture for SIP registration in IMS

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

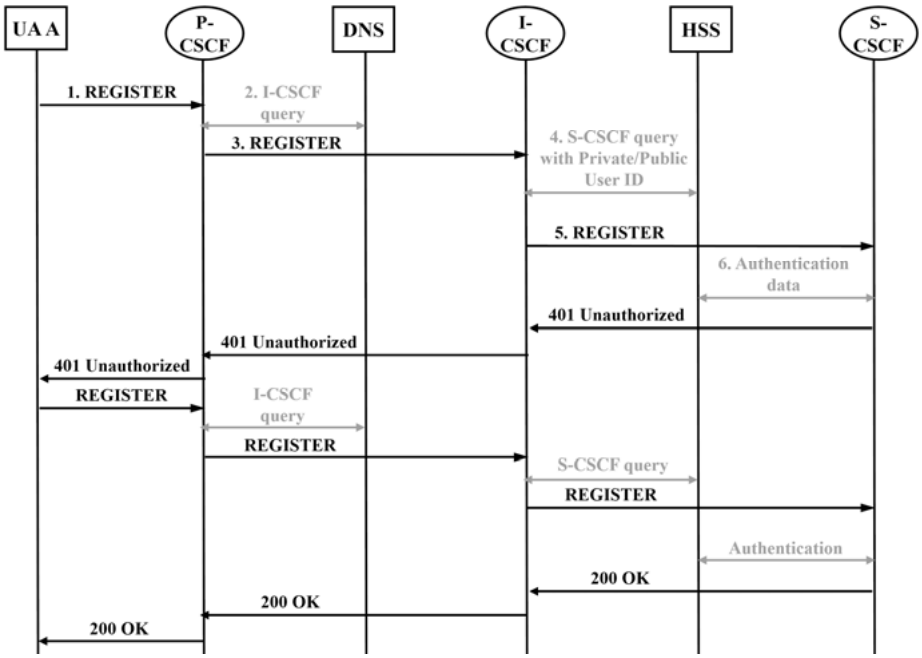

Fig. 2.5: Procedure of a SIP registration in the IMS

Figure 2.6 then shows the network architecture with the focus for 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.

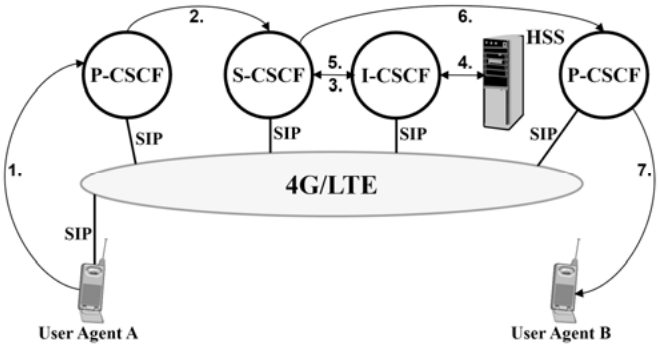

Fig. 2.6: Network architecture for a SIP session setup in the IMS

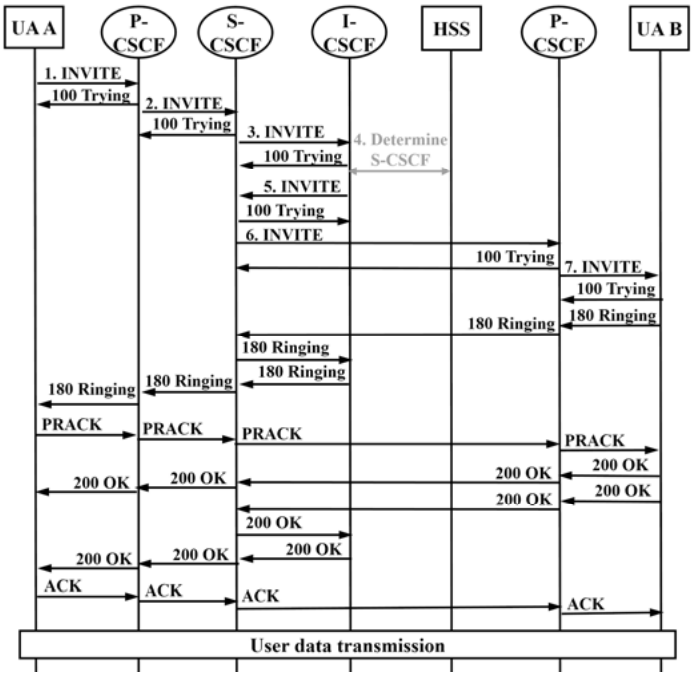

Fig. 2.7: Procedure for establishing a SIP session in the IMS

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].

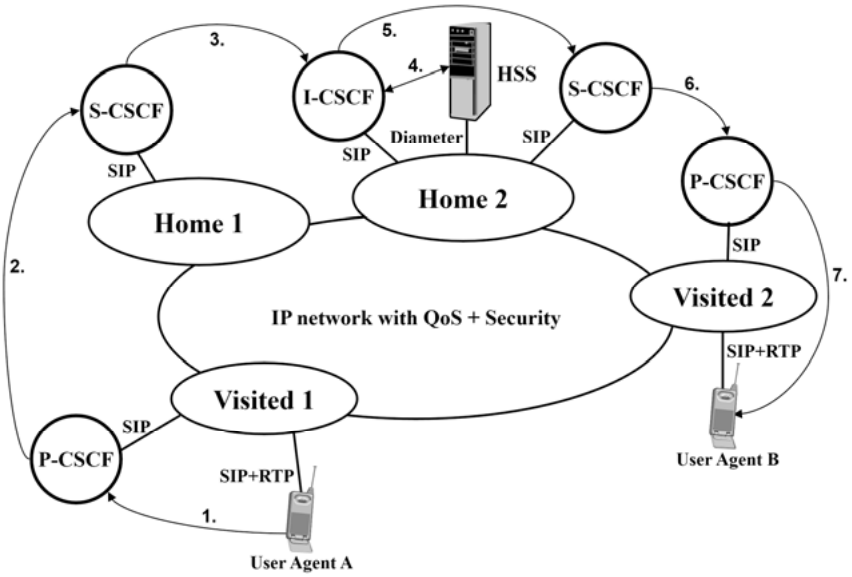

Fig. 2.8: Network architecture for roaming in IMS-based networks

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 entire 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 PSTN (SCN bearer channel, switched circuit network). The X represents the communication relationship between the terminations. Context 3 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].

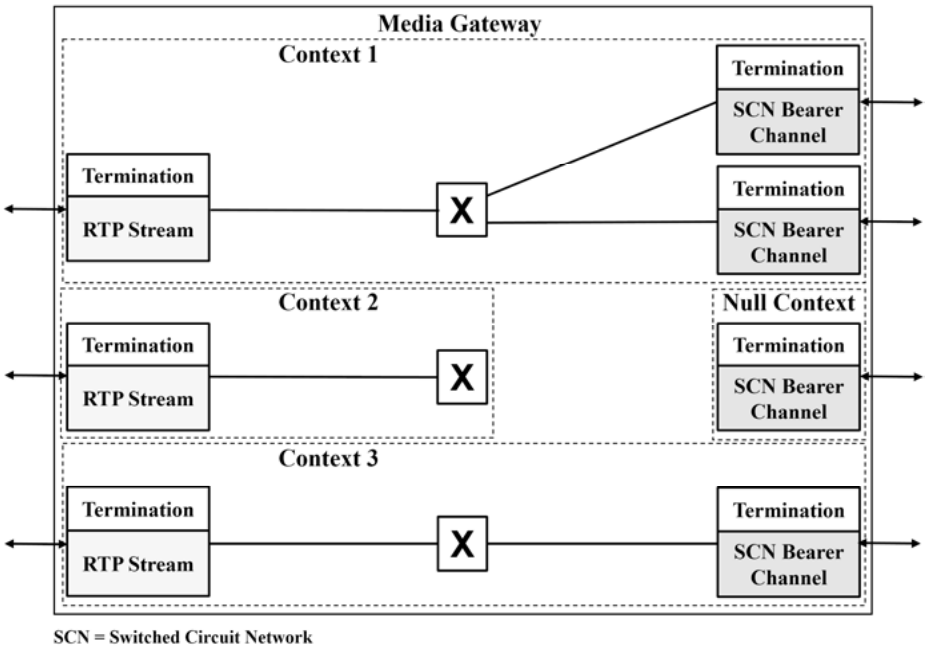

Fig. 2.9: Example of an H.248 connection model with contexts [104; 162]

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. Ee.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.

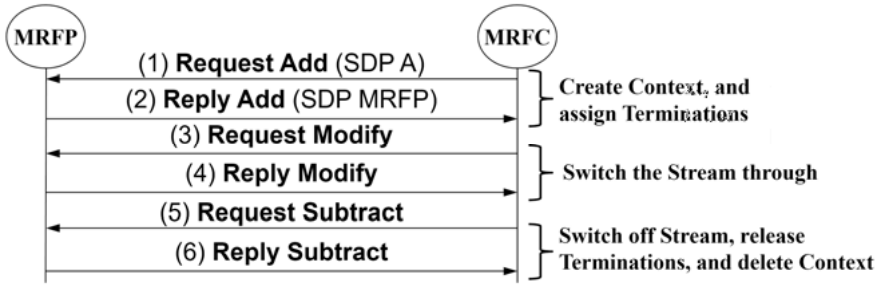

Fig. 2.11: Exemplary H.248 message flow for providing a media stream for user A

```
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): fmtp:116 mode-change-capability=2; max-red=0
               Media Attribute (a): fmtp:118 mode-change-capability=2; max-red=0
               Media Attribute (a): fmtp:111 0-15
               Media Attribute (a): fmtp:110 0-15
               Media Attribute (a): ptime: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
v Command: Add
     Command: Add
     Termination ID: rtp/1/1016919
   v Descriptors
     v Media Descriptor
          StreamID: 1
       v Local Descriptor
         v 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): fmtp:116 max-red=0; mode-change-capability=2
            > Media Attribute (a): fmtp:111 0-15
            > Media Attribute (a): ptime: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
v Command: Priority
     Command: Priority
     Priority: 9
v Command: Modify
     Command: Modify
     Termination ID: rtp/1/1016919
   v Descriptors
     v Signal Descriptor
       v 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].

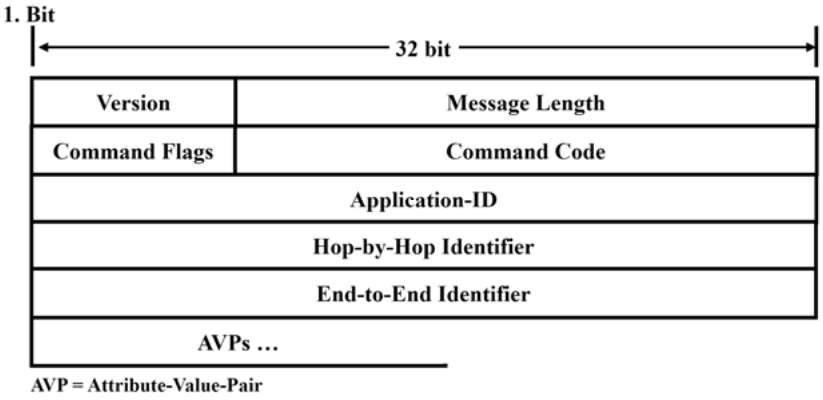

Fig. 2.16: Header of a Diameter message [12]

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].

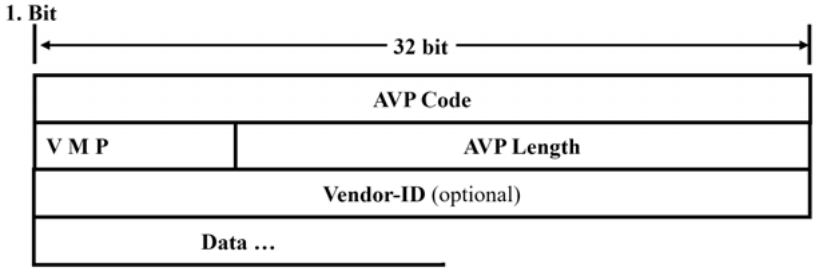

Fig. 2.17: AVP header [12]

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

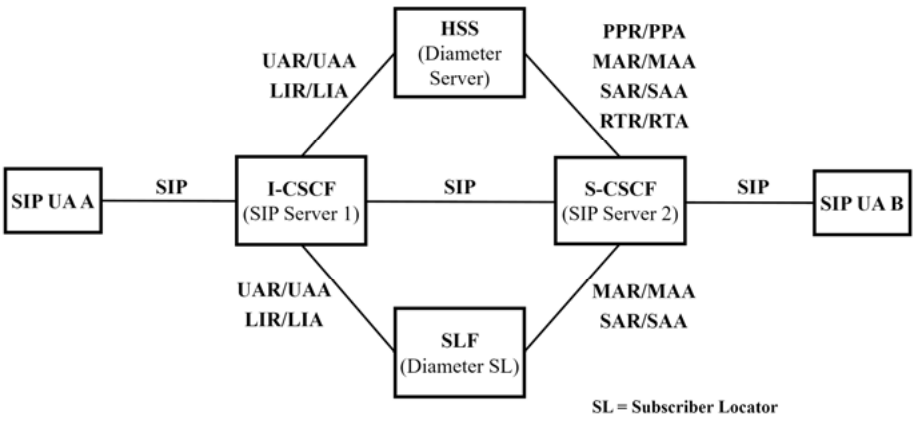

Fig. 2.18: SIP network architecture and Diameter protocol [11]

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:scscf.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 her/his 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:scscf.mnc001.mcc001.3gppnetwork.org:5060) with Location-Info-Answer ((16) LIA, see Figure 2.27) in the AVP server URI or server name.

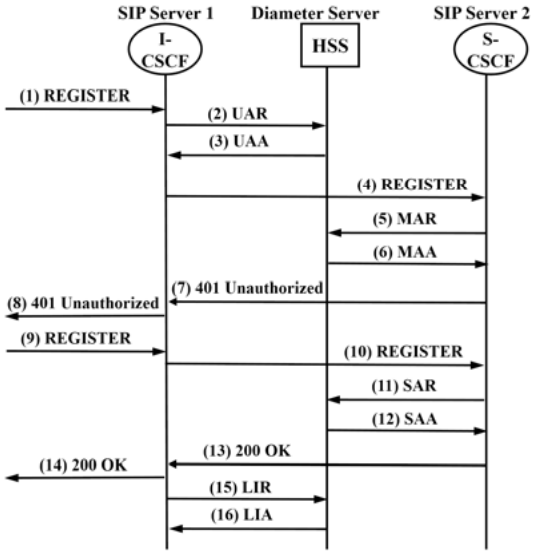

Fig. 2.19: SIP authentication by Diameter at registration

```
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. = 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: 0x4ccdf17c
   [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: 000000
 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: 0000025dc0000010000028af000000010000025ac0000038...
      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

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

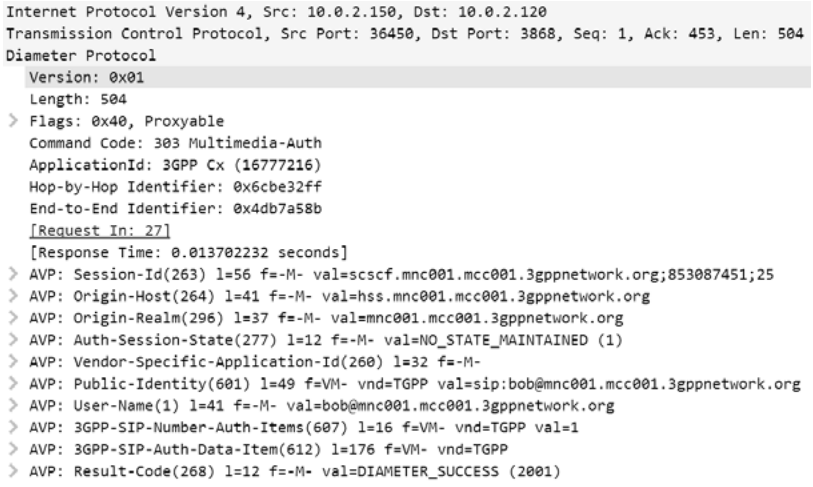

Fig. 2.23: MAA Diameter message (6) captured with protocol analysis software

```
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
 v 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)
 v 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
v 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...
   v eXtensible Markup Language
     > <?xml
     v <IMSSubscription>
       > <PrivateID>
       v <ServiceProfile>
         > <PublicIdentity>
         v <InitialFilterCriteria>
           > <Priority>
           > <TriggerPoint>
           v <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: 0x3d493cbb
   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: 0x3d493cbb
   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-
 v 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].

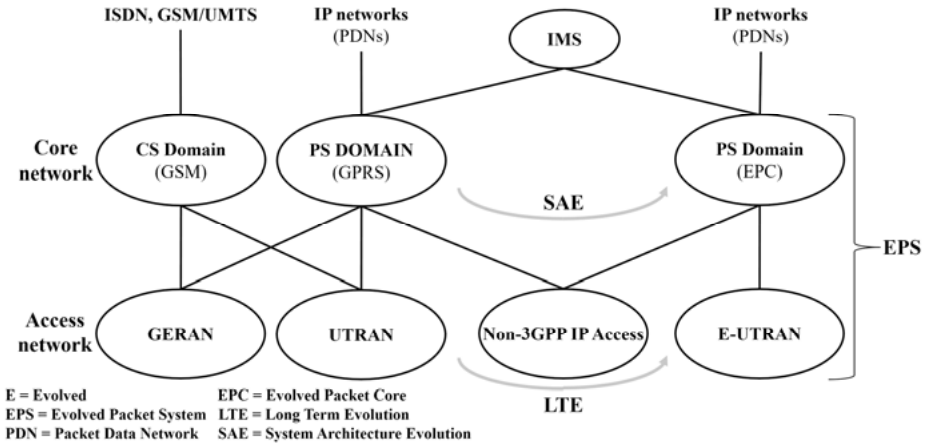

Fig. 2.28: Transformations in 3GPP networks towards SAE and LTE [173]

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 UTRAN, E-UTRAN, 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 will be switched off in the long 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.

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].

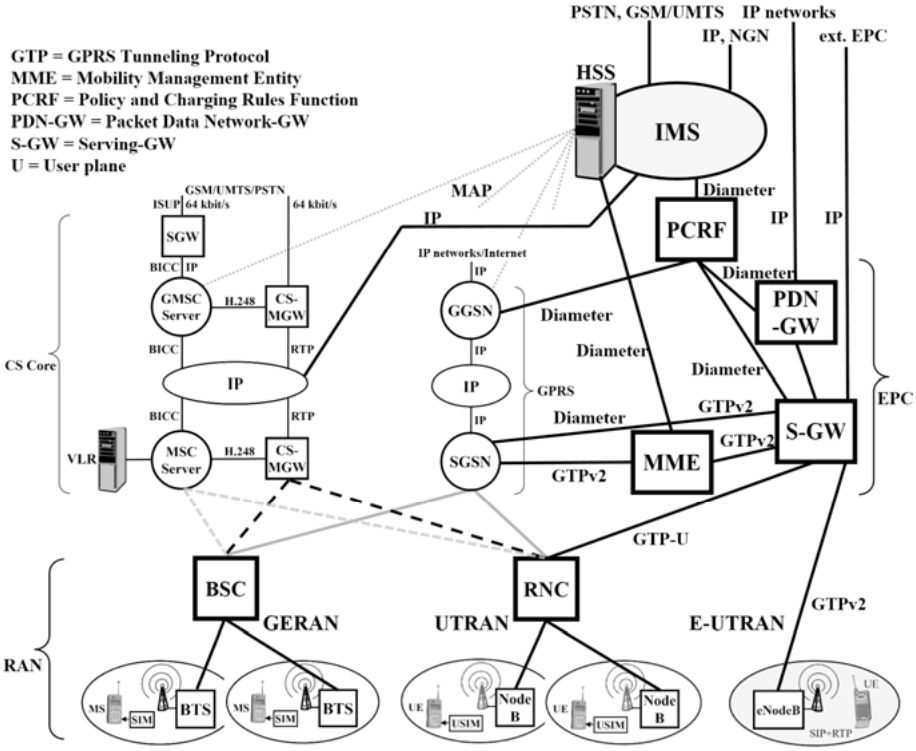

Fig. 2.29: 3GPP Release 8 mobile network

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 Fallback (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].

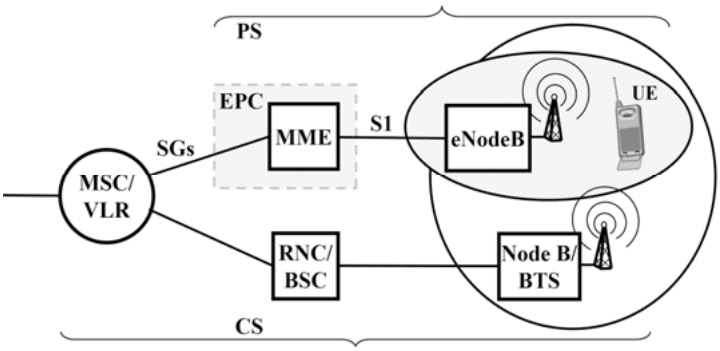

Fig. 2.30: Mobile network with LTE and Circuit Switched Fallback (CSFB) for telephony

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.

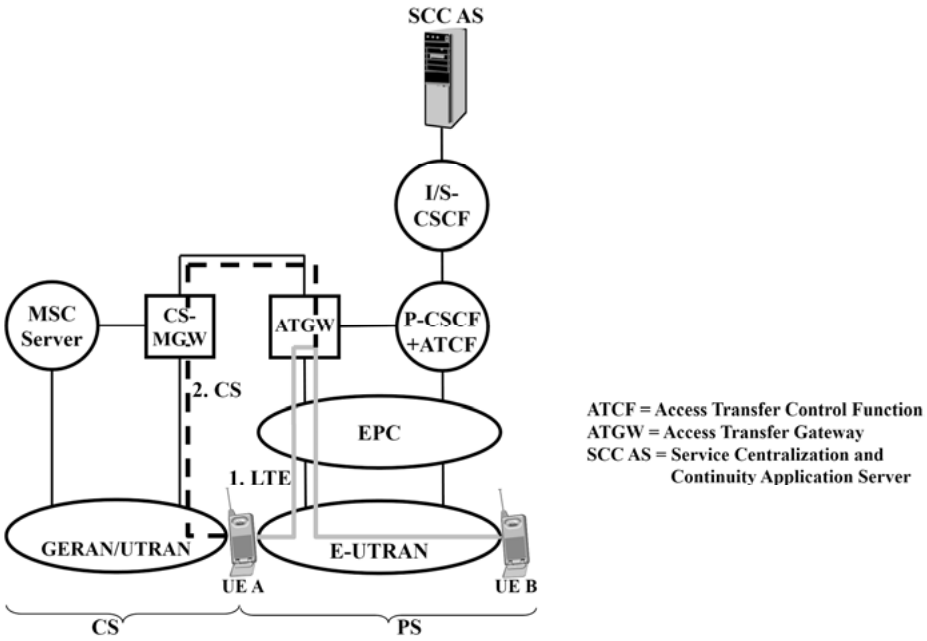

Fig. 2.31: VoLTE with SRVCC

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. A handover from the CS to the PS network is not supported [158; 34].

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 hando-

ver, 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].

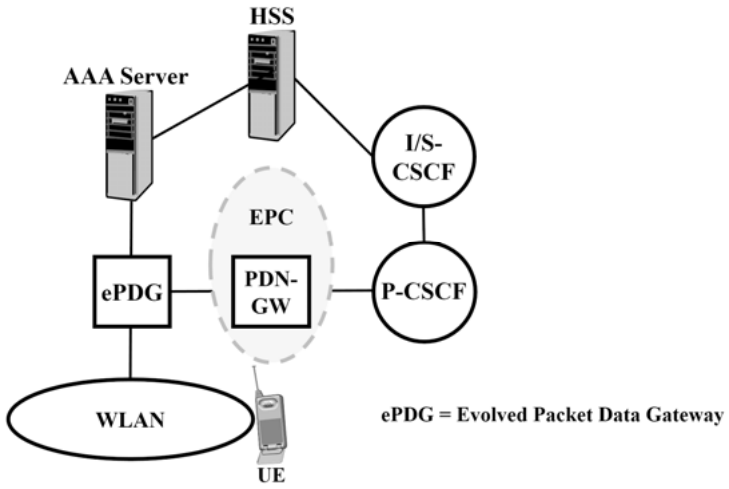

Fig. 2.32: VoWifi in 3GPP mobile network since Release 8

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 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 is still common 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].

https://doi.org/10.1515/9783110724509-003

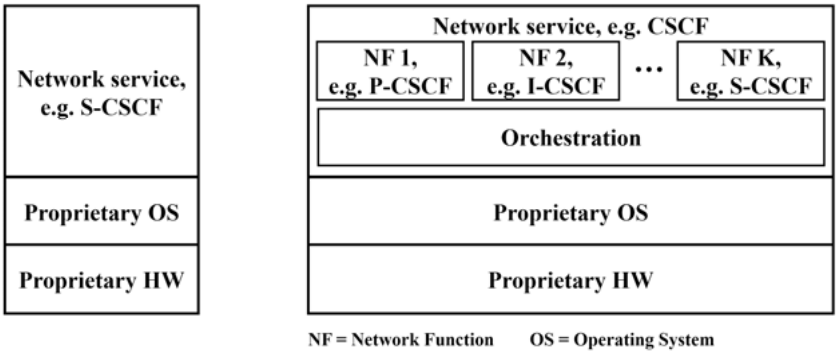

Fig. 3.1: Implementation of network elements or network services with proprietary hardware

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 has 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 in 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].

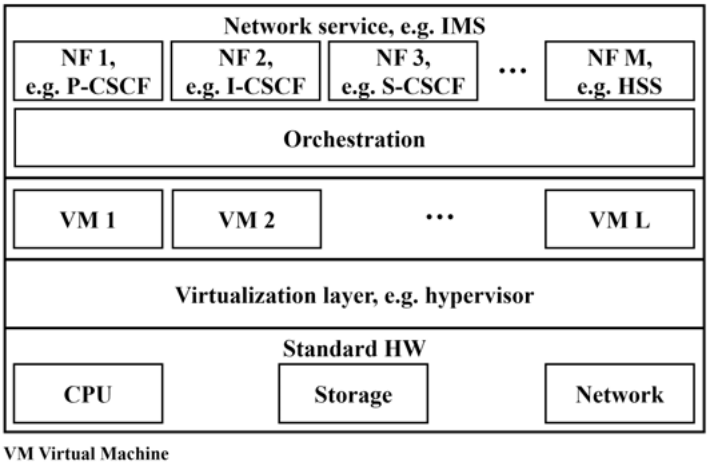

Fig. 3.2: NFV-based implementation of network elements or network services with standard HW using the example of IMS

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 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.

Figure 3.3 [141] gives a first overview of the NFV framework, according to ETSI. 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].

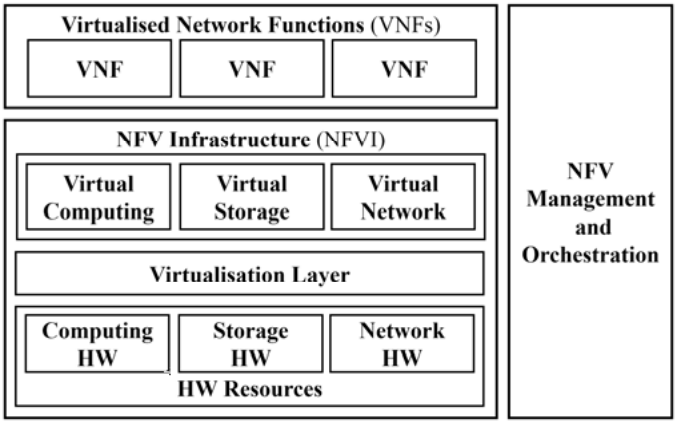

Fig. 3.3: Overview of the NFV framework [141]

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 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., instantiation (creating a VNF), up-

date/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].

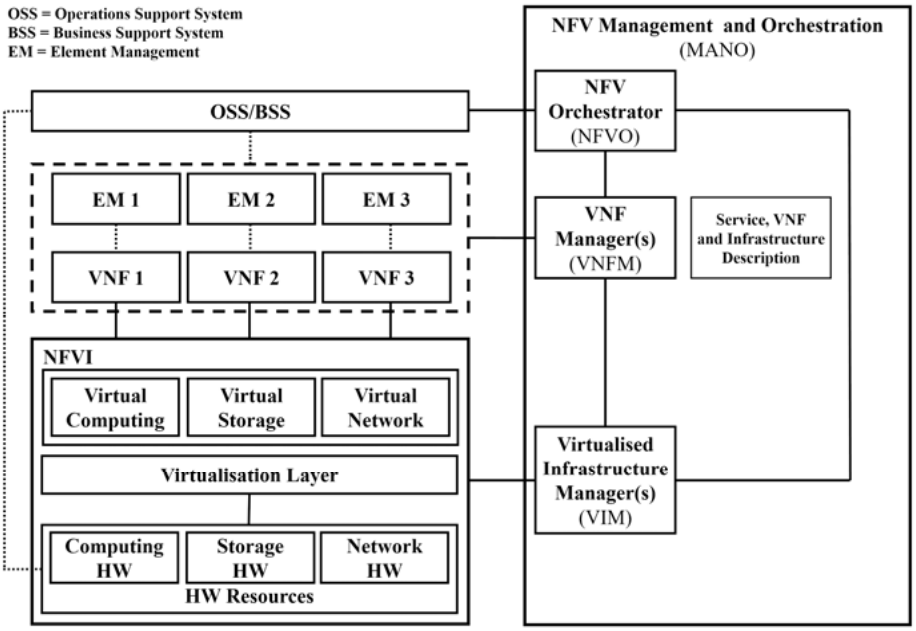

Fig. 3.4: NFV reference architecture framework according to ETSI [141; 132]

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 & 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].

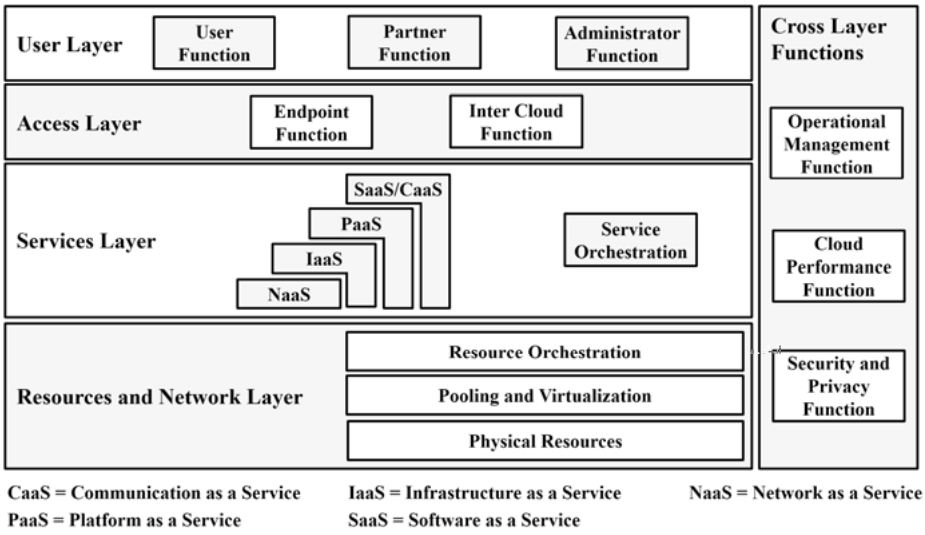

Fig. 3.5: Functional cloud computing reference architecture [100]

[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.

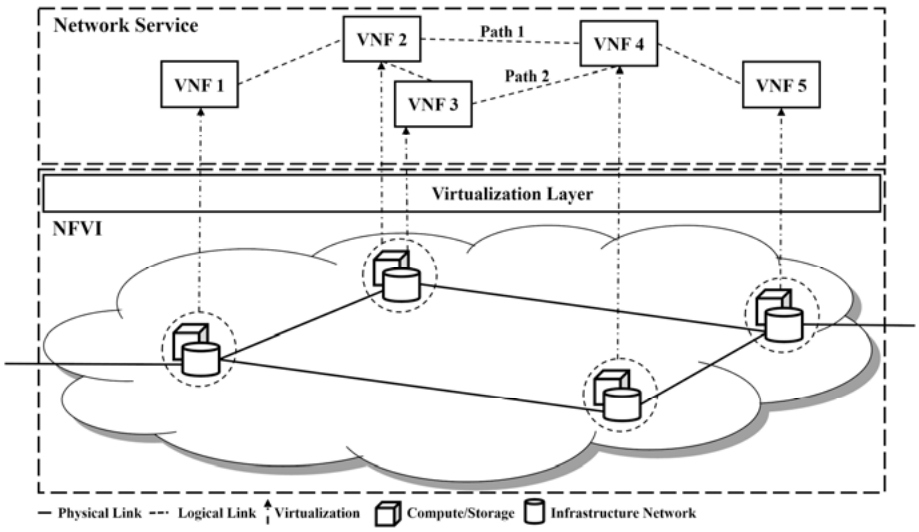

Fig. 3.6: Example of VNF-FG-based network services with VNFs in operation at different locations [132]

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 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, 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].

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 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; 94].

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 3.7 gives an overview. A so-called MEC host provides the 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].

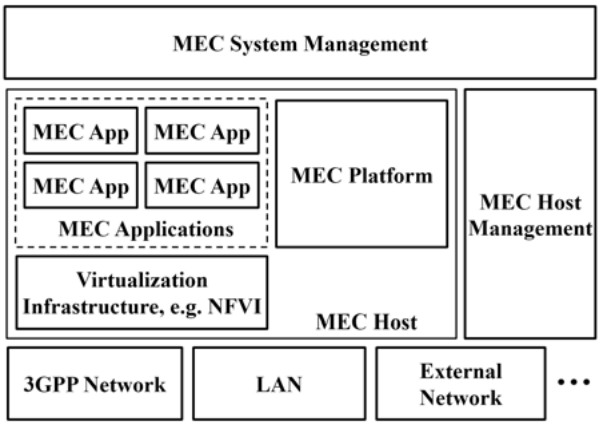

Fig. 3.7: Multi-access Edge Computing framework [137]

Figure 3.8 shows the MEC reference architecture, according to [137] in more detail. 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. NFV and MEC, therefore, complement each other and also use the same techniques.

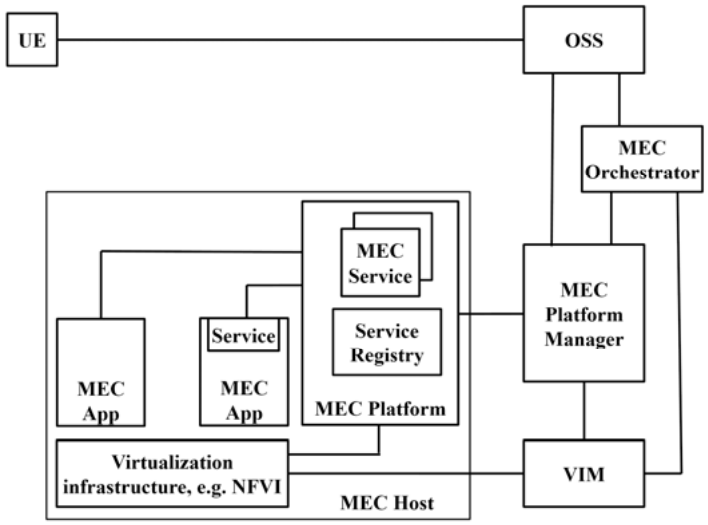

Fig. 3.8: MEC reference architecture [137]

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 an HW server 1 to an 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 tunneling 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, which 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 to be 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 3.9. This is done by introducing a central SDN controller for the control plane and decentralized, pure SDN switches reduced to data packet forwarding [173].

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 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 3.9 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 central 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].

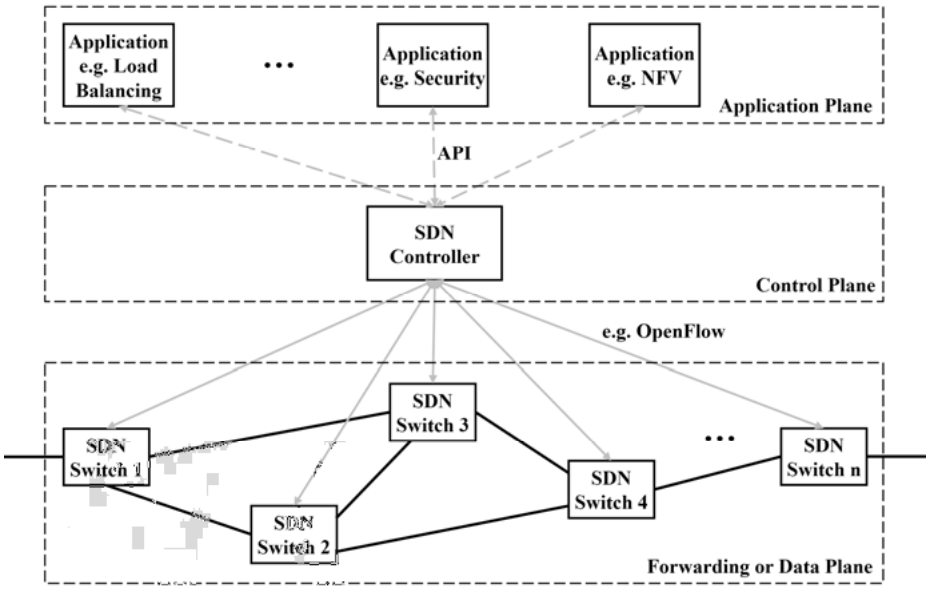

Fig. 3.9: SDN architecture

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 3.10 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 parame-

ters 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].

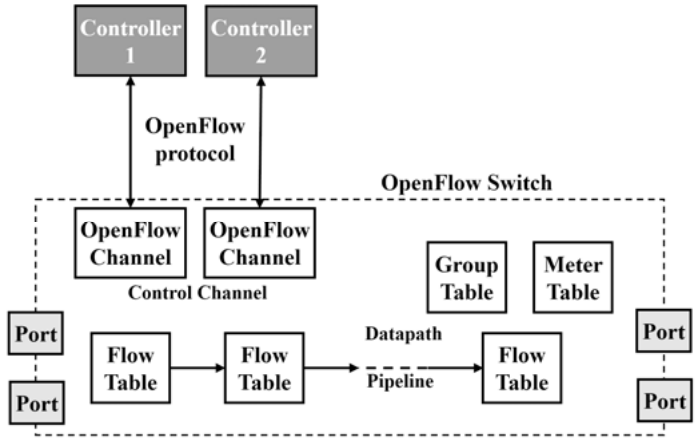

Fig. 3.10: OpenFlow switch [149]

The functionality of a flow table can be described even more concretely using Figure 3.11. 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.

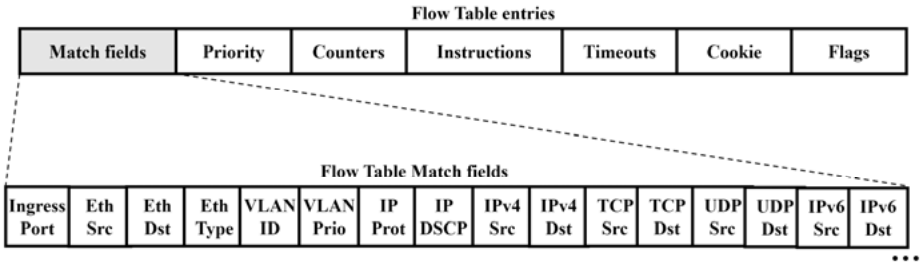

Fig. 3.11: Flow table [166]

The match fields can contain, among others, the fields shown in Figure 3.11. 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 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 stat 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 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 3.12, 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].

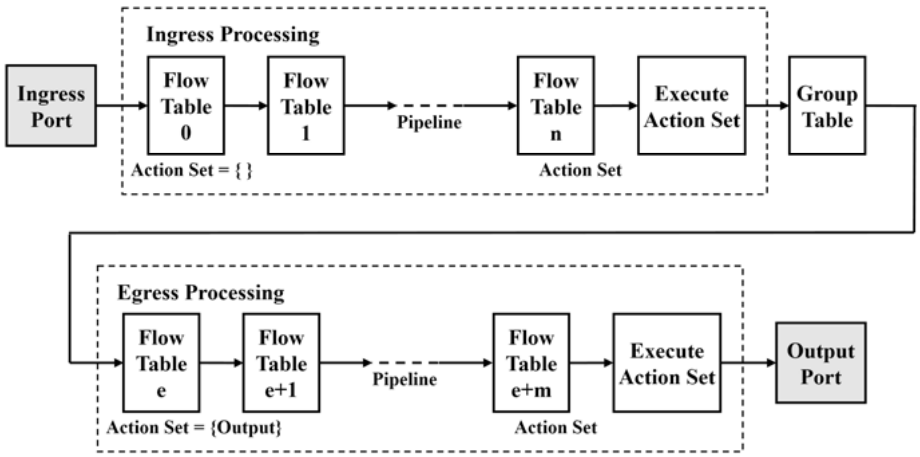

Fig. 3.12: Packet processing in the pipeline [149]

Group tables with the structure shown in Figure 3.13 can also be involved in the processing. It provides so-called action buckets for a whole group of flows, even taking into account several ports, whereby one action set per action bucket is stored for execution [149].

Group Identifier	Group Type	Counters	Action Buckets

Fig. 3.13: Group table [149]

Figure 3.14 shows the processing of a packet received by an OpenFlow Switch.

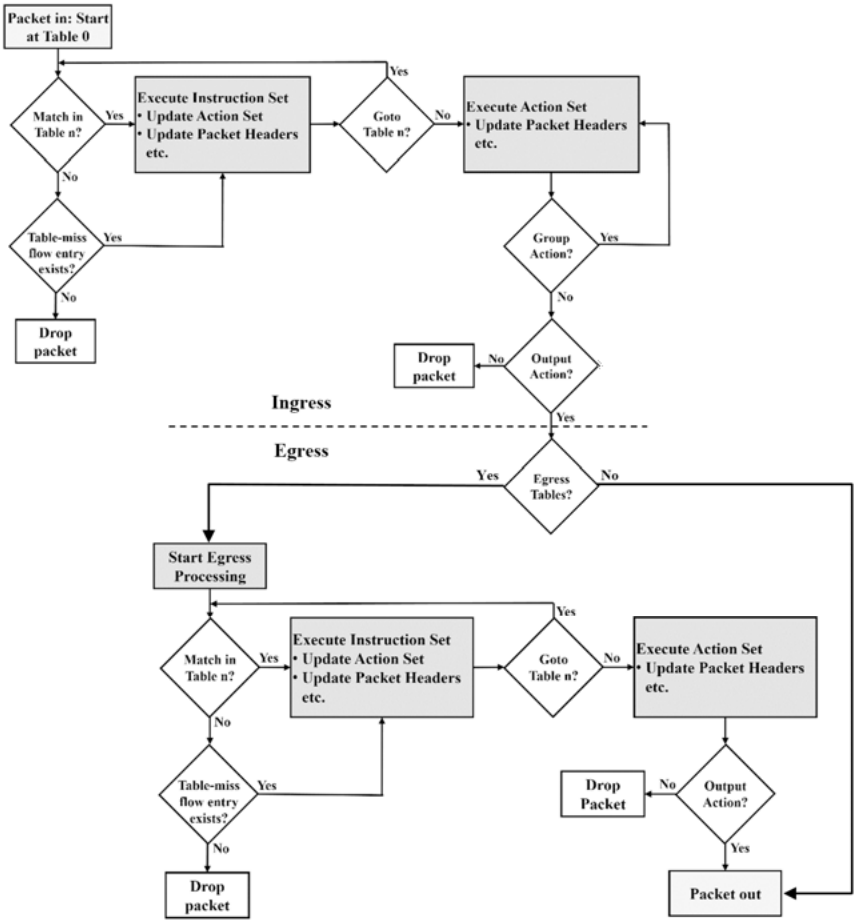

Fig. 3.14: Packet processing in OpenFlow switch [149]

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 3.14 [149; 166].

Figure 3.15 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 *.

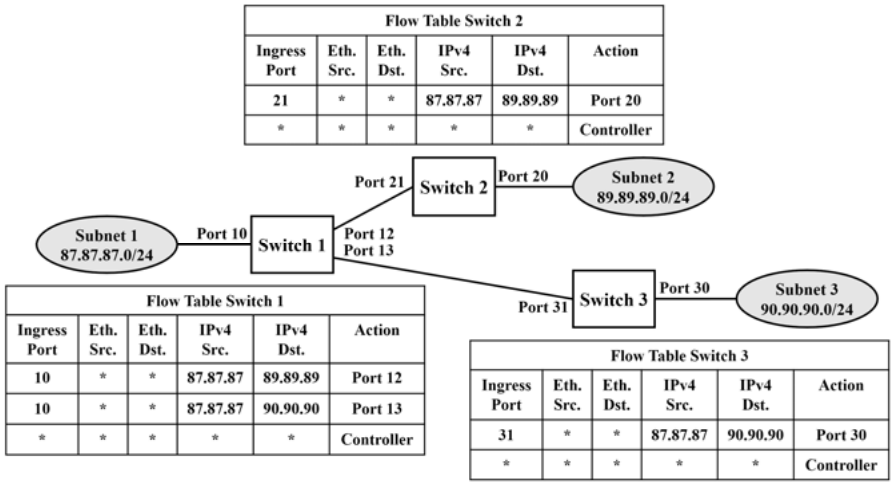

Fig. 3.15: Example for IP routing with SDN switches

According to Figure 3.16 [102], the communication between 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, symmet-

rical 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.

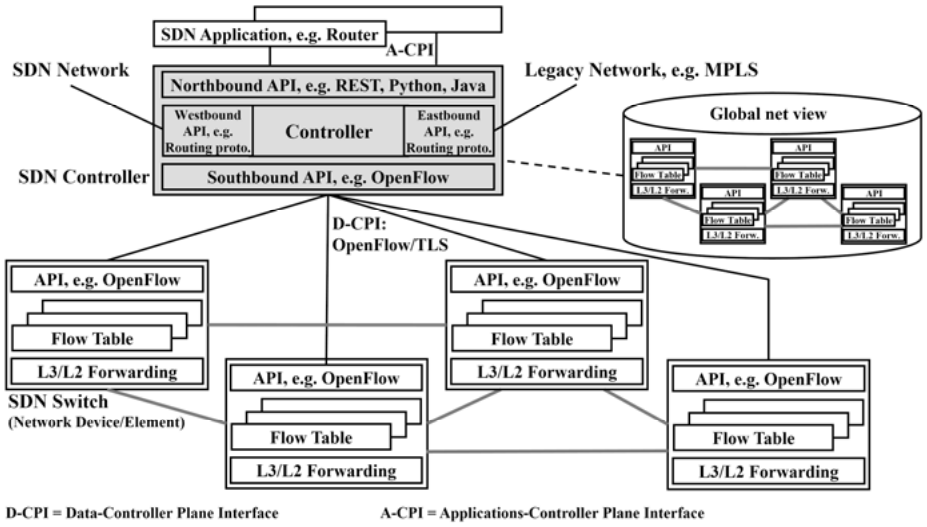

Fig. 3.16: SDN controllers and SDN switches in the network

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 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 specific OpenFlow process was successfully completed.

Message	Function
Role-Request	For setting or querying the OpenFlow channel or the controller ID in a switch, 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
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 it 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 3.17 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].

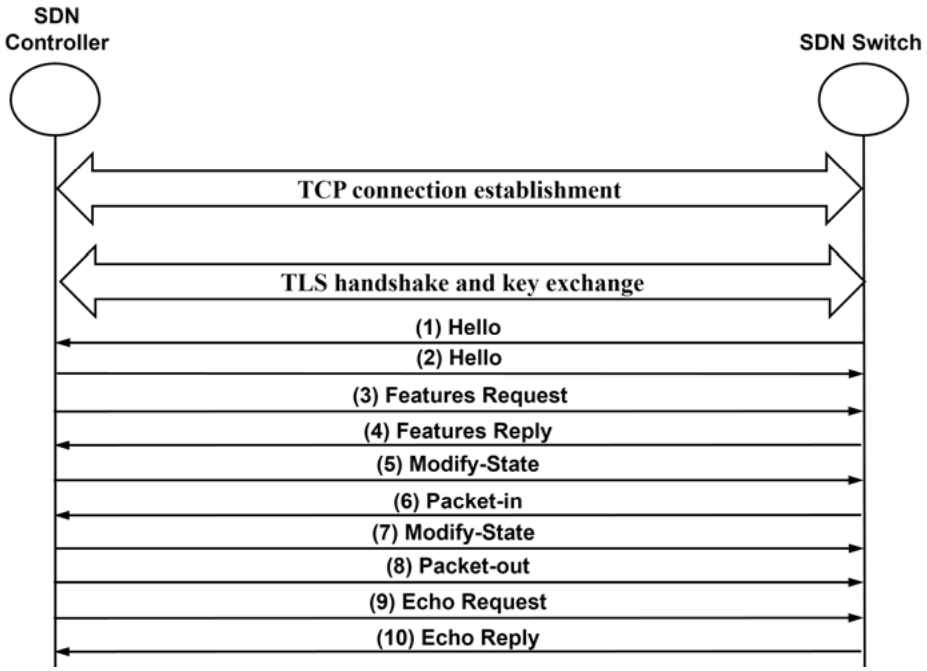

Fig. 3.17: Exemplary OpenFlow message exchange

In addition to the OpenFlow message sequence chart in Figure 3.17, the following Figures 3.18 to 3.22 describe the structure of OpenFlow messages in more detail. Figure 3.18 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 3.19 shows a first practical example with a Hello message ((1) in Figure 3.17) 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 3.17, Type OFPT_FEATURES_REPLY), according to Figure 3.20 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.

Figure 3.21 represents a Packet-in OpenFlow message ((6) in Figure 3.17, 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 3.22 ((7) in Figure 3.17, 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).

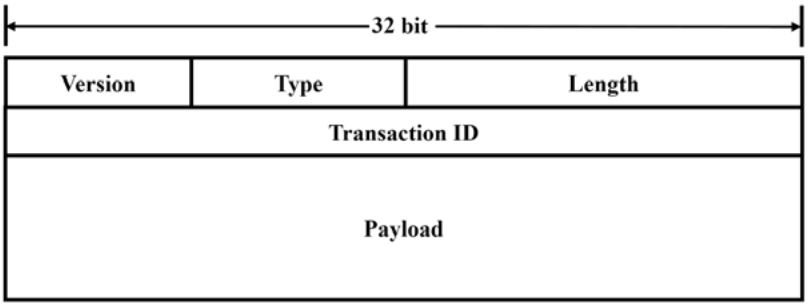

Fig. 3.18: Structure of the header of an OpenFlow message

```
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. 3.19: 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. 3.20: Features Reply OpenFlow message (4) captured with protocol analysis software

```
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. 3.21: 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
 v Match
      Type: OFPMT_OXM (1)
      Length: 22
   v OXM field
         Class: OFPXMC_OPENFLOW_BASIC (0x8000)
         0000 000. = Field: OFPXMT_OFB_IN_PORT (0)
         .... ...0 = Has mask: False
         Length: 4
         Value: 2
   v 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
 v Instruction
      Type: OFPIT_APPLY_ACTIONS (4)
      Length: 24
      Pad: 00000000
   v Action
         Type: OFPAT_OUTPUT (0)
         Length: 16
         Port: 1
         Max length: 65509
         Pad: 000000000000
```

Fig. 3.22: 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].

It is interesting to note that although there are several specifications for the southbound API, the D-CPI in Figure 3.16, 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 3.16, 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 also 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 by a VIM in the NFV MANO, as shown in Figure 3.23 [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 3.23, can be found in [97].

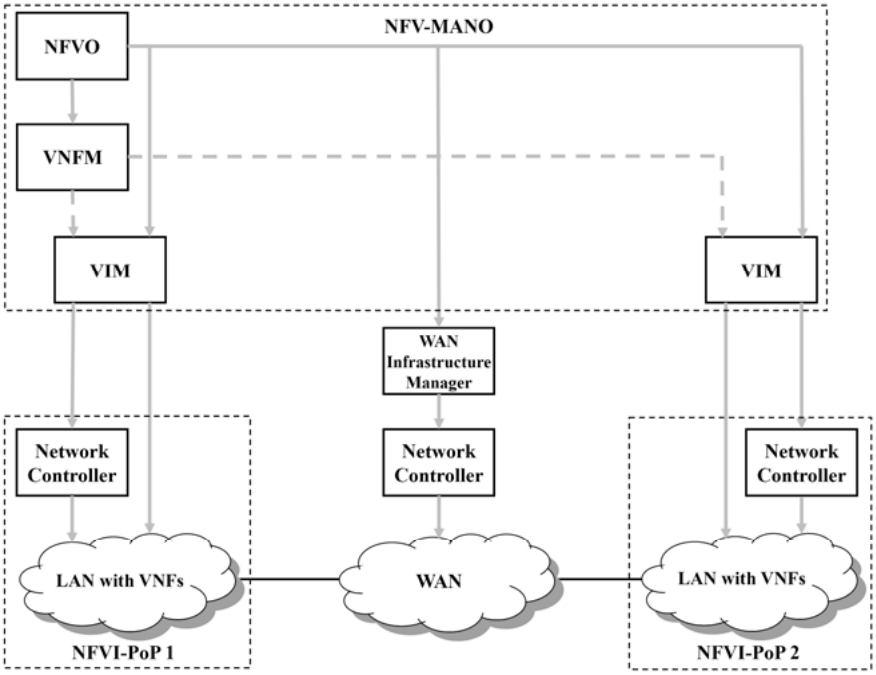

Fig. 3.23: NFV MANO and SDN controller [132]

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 3.24 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 3.24

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].

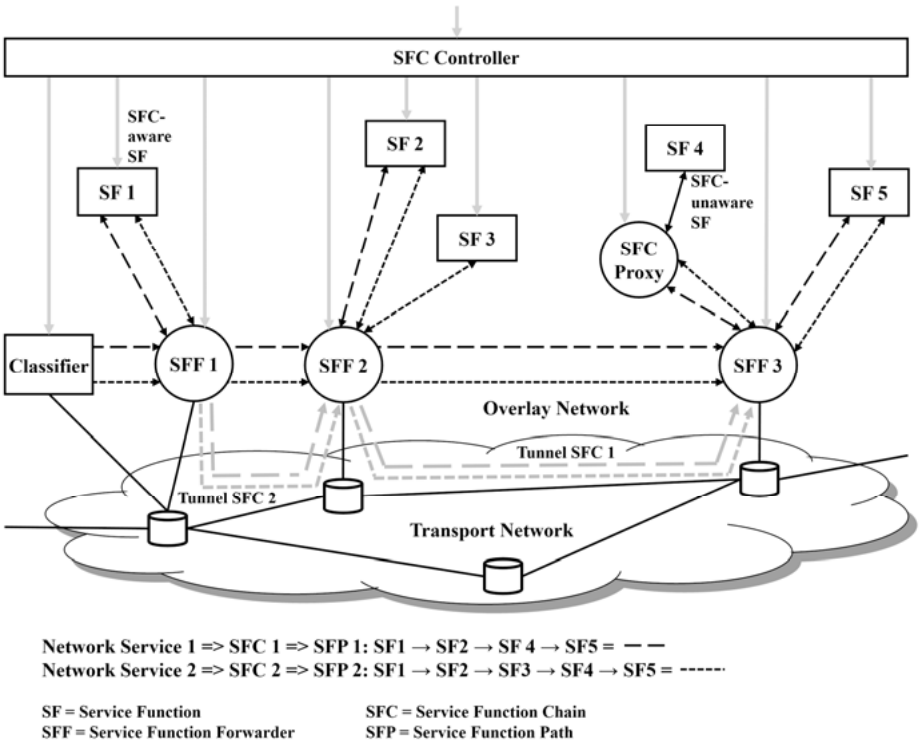

Fig. 3.24: SFC architecture

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 3.25 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 3.26 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 3.24 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].

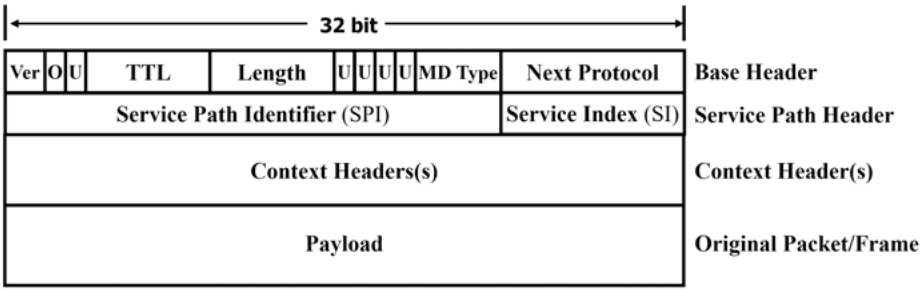

Fig. 3.25: Structure of the Network Service Header (NSH)

```
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 (0x009876)
   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. 3.26: Encapsulated Ethernet frame with NSH header captured with protocol analysis software

In the transport network shown in Figure 3.24, 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 3.24 can be carried out as for SDN in Figure 3.23: 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 FNs in terms of four objectives and twelve corresponding design goals, which are explained briefly in the following and are shown in Figure 3.27.

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

- Service awareness: FNs should provide current and future services that are 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: FNs 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: FNs 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: FNs should take social and economic issues into account in such a way 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 already played a significant role in the NGN concept [173].

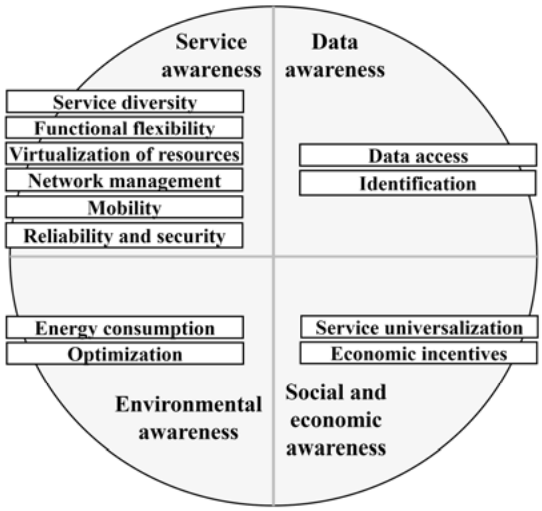

Fig. 3.27: Four objectives and twelve design goals for Future Networks [179]

As already mentioned above and shown in Figure 3.27, 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 3.28 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 in turn 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].

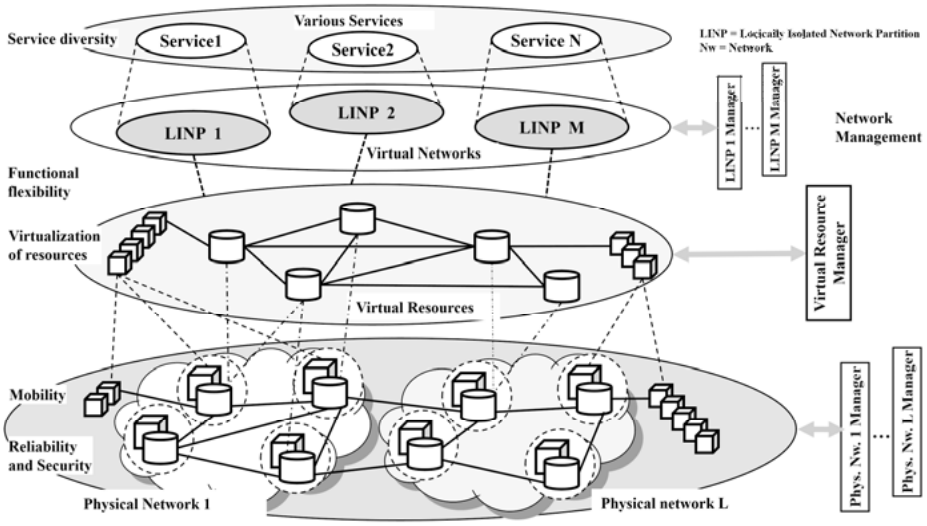

Fig. 3.28: Virtualization in Future Networks [180]

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 3.28 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. Of which 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 basic 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 [188] and the TIA (Telecommunications Industry Association) [192], in China, the IMT-2020 Promotion Group [189], in South Korea, the 5G Forum [190], and in Japan, the 5GMF (Fifth Generation Mobile Communications Promotion Forum) [191]. These efforts have been and are being supplemented by many research projects on 5g at universities, in companies, and joint projects

4.1 5G Use Cases and Usage Scenarios

The METIS project published scenarios, use cases (here called test cases), and 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 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

https://doi.org/10.1515/9783110724509-004

– 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]

	Scenario				
	Amazingly fast	Great service in a crowd	Ubiquitous things communicating	Best experience follows you	Super real-time and reliable connections
	Essential requirements				
	Very high data rate	Very dense crowds of users	Very low energy, cost, and a massive number of devices	Mobility	Very low latency
Use/test case					
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 computing for mobile terminals				X	X
Open-air festival		X	X		
Emergency communications			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 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]

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

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 and remote surgery	≤ 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 km^2 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 summarised in three main areas:

- Enhanced Mobile Broadband
- Ultra-Reliable and Low Latency Communications
- Massive Machine Type Communications.

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.

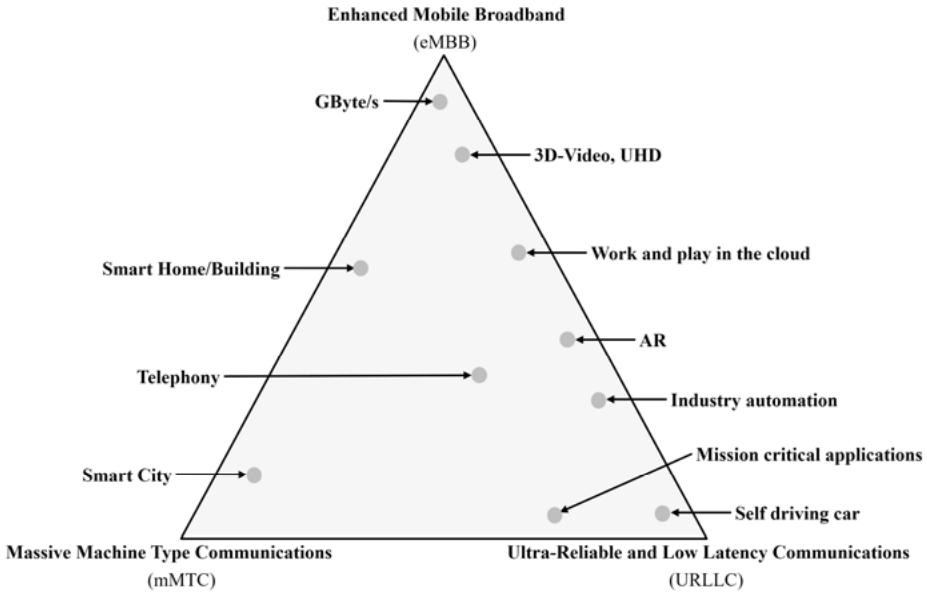

Fig. 4.1: ITU-R usage scenarios of IMT-2020 [128]

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 maintains 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 are 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 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 dealt with in Section 4.3, in particular, based on further standards.

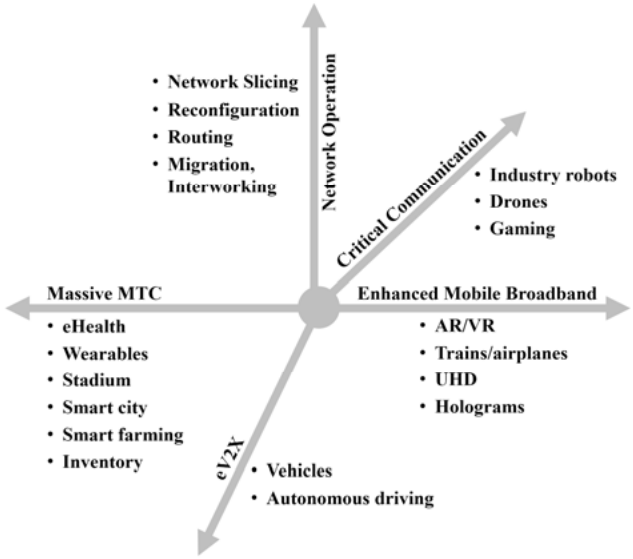

Fig. 4.2: 3GPP usage categories and examples of use cases for 5G [27]

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

Category	General requirements
eMBB (Enhanced Mobile Broadband)	– Very high data rate, up to 10 Gbit/s per user – Low latency – High traffic density, Tbit/s/km^2 – High density for UE, up to 2500 UEs/km^2 – 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)	– 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)/km^2 – Precise position, ≤ 10 cm
MIoT (Massive Internet of Things) or mMTC (Massive Machine Type Communications)	– 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)/km^2 – High positioning accuracy, ≤ 50 cm

Category	General requirements
eV2X (Enhancement of Vehicle-to-Everything)	– 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 will be used in many areas of life. [88] provides an incomplete but illustrative overview of applications and industries that are 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, 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 are being 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 are 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 will be available across WANs in the production process end-to-end.

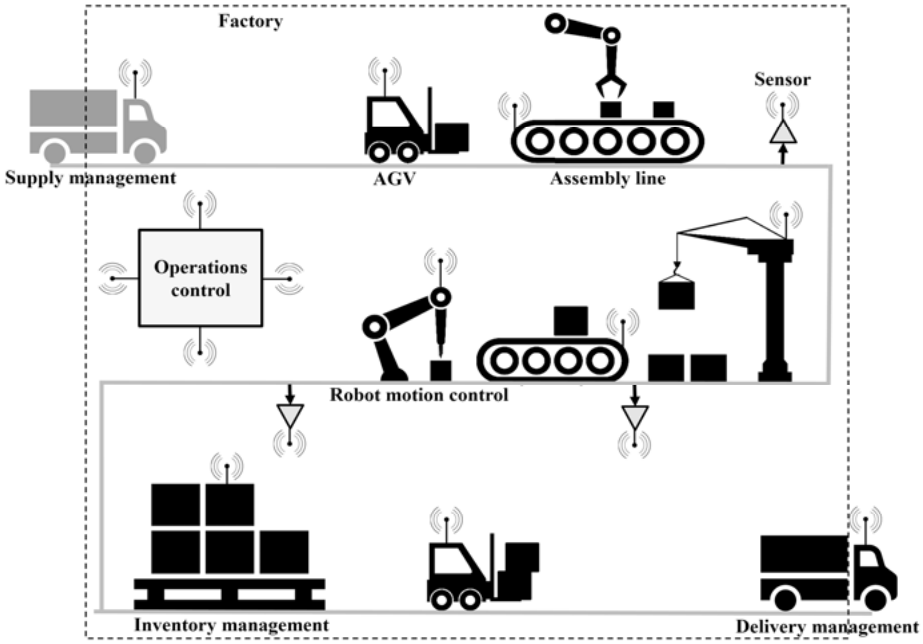

Fig. 4.3: Exemplary application areas of a 5G network in the factory of the future [71]

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).

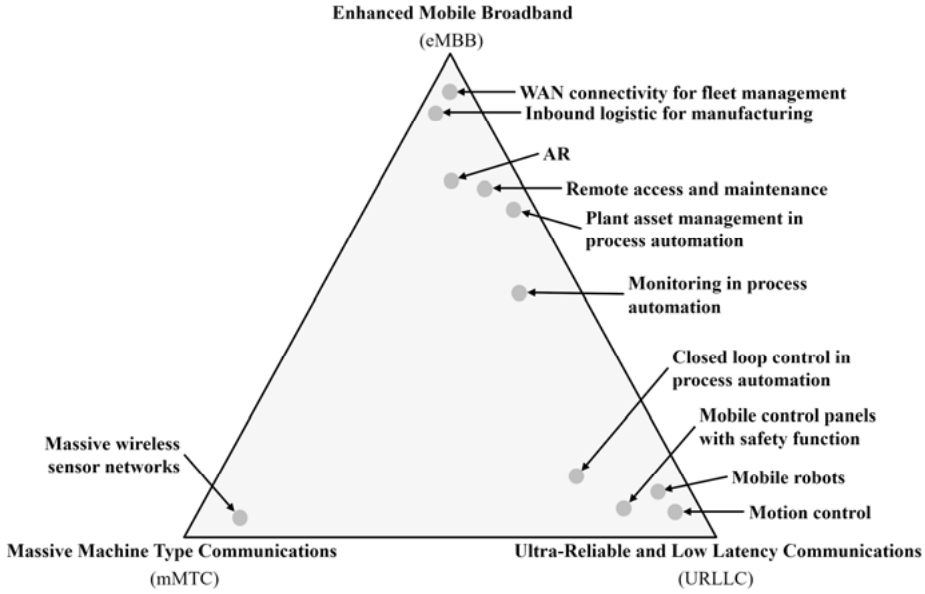

Fig. 4.4: 5G-ACIA use cases and ITU-R/3GPP usage categories [71]

The 5GAA (5G Automotive Association) is 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	– Emergency braking – Intersection management assist – Collision warning – Lane change
Vehicle operations management	– Sensors monitoring – Software updates – Remote support
Convenience	– Infotainment – Assisted and cooperative navigation – Autonomous smart parking
Autonomous driving	– Control if autonomous driving is allowed – Tele-operation (potentially with AR support for a remote driver) – Handling of dynamic maps (update/download)
Platooning	– 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	– Greenlight optimal speed advisory – Traffic jam information – Routing advise, e.g., smart routing
Society and community	– 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 which then 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].

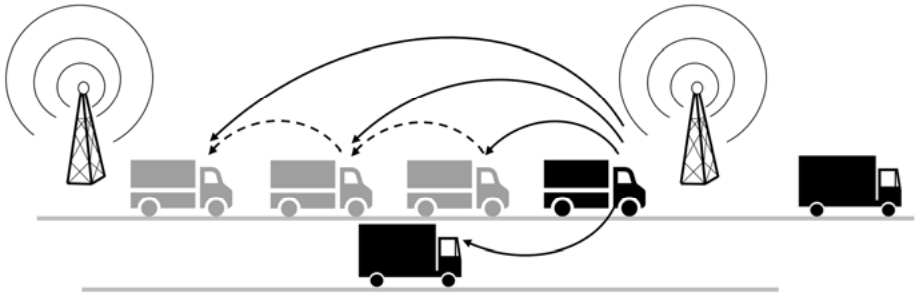

Fig. 4.5: Platooning with trucks

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 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 will 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 that an official standardization 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/km^2
- 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/m^2.

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].

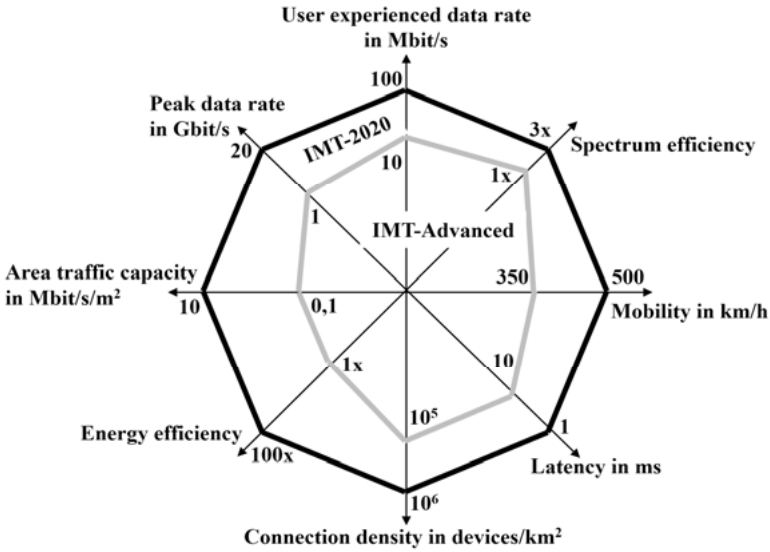

Fig. 4.6: Key capabilities of an IMT-2020 system (5G) in comparison with IMT-Advanced (4G) [128]

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.

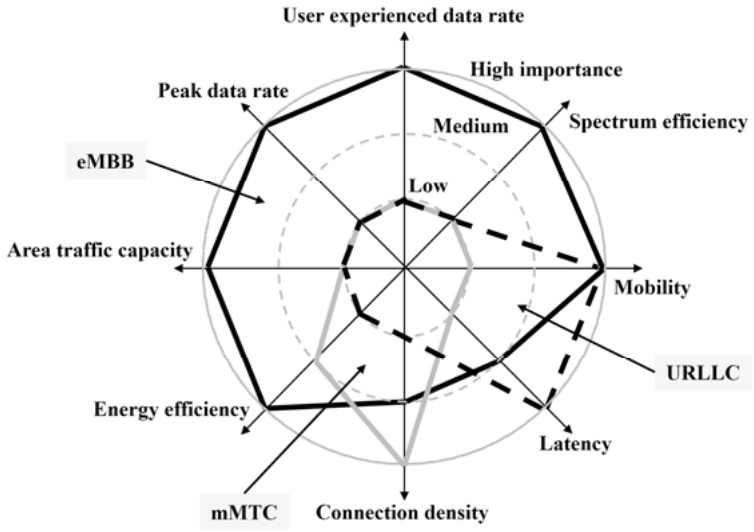

Fig. 4.7: IMT-2020 key capabilities in usage scenarios eMBB, URLLC, and mMTC [128]

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. In terms of requirements, the most crucial 3GPP specification is TS 22.261, which is available in three versions according to the above releases. All three documents divide 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 Release 16 that is currently in progress. 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 (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 8.3) 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 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, or 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/km^2
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
Broad-band 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
Broad-cast-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/km²	100000/km²	1000m x 1000m x 30m
Process automation – remote control	60 ms	99,999	99,999	1-100 Mbit/s	100 Gbit/s/km²	1000/km²	300m x 300m x 50m
Process automation – monitoring	60 ms	99,9	99,9	1 Mbit/s	10 Gbit/s/km²	10000/km²	300m x 300m x 50m
Electricity distribution – medium voltage	40 ms	99,9	99,9	10 Mbit/s	10 Gbit/s/km²	1000/km²	100 km along power line
Electricity distribution – high voltage	5 ms	99,999	99,999	10 Mbit/s	100 Gbit/s/km²	1000/km²	200 km along power line
Intelligent transport systems – infrastructure backhaul	30 ms	99,999	99,999	10 Mbit/s	10 Gbit/s/km²	1000/km²	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 will provide enhancements and also completely new service requirements and features. Table 4.9 summarizes the most important service requirements for Release 16 [22].

Tab. 4.9: Additional service requirements for a 5G system in 3GPP Release 16 [22]

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 5G access via satellite
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) und high connection density (1000/km^2)
Positioning accuracy absolute up to 0.3 m horizontal and 2 m vertical when moving the UE at up to 60 km/h, or relative 0.2 m horizontal, 0.2 m vertical, at up to 30 km/h
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

Security
For Satellite RAN, 5G-LAN
Enhancements regarding data integrity and encryption
Charging aspects
For Satellite RAN, 5G-LAN

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; 22]. However, this does not mean that all the resulting necessary features and functions will be standardized in the same release. An example of this is the satellite-based RAN mentioned in Table 4.9 for Release 16. The result for this in Release 16 is only one study in the form of TR 22.822 [25]. The features and functions to be standardized for the technical implementation will be the subject of Release 17 (see Section 12.1).

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 the still incomplete TR 21.916 [20].

5 5G Standardization and Regulation

As mentioned in Chapter 4, the ITU has specified a 5G target system under the IMT-2020 designation. 3GPP 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, will be 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 will be 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 [47] 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

https://doi.org/10.1515/9783110724509-005

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. Here, however, there are only short radio ranges and substantial impairments by obstacles (e.g., water vapor, fog, rain, leaves, also people) in the radio path [127]. Accordingly, the radio transmission and antenna technology must become more complex. Also, an application is locally limited.

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 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, in the near future 6 GHz) or LPWAN (Low Power Wide Area Network, above 433 or 863 or 902 MHz) are also taken into account.

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

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

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

NR frequency band	UL	DL	Duplex mode
n257	26500 MHz – 43000 MHz	26500 MHz – 43000 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

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 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. Here a 5G target system was specified 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 Co-ordination Group (PCG)		
TSG (Technical Specification Group) **RAN** • **Radio Access Network**	**TSG SA** • **Service & Systems Aspects**	**TSG CT** • **Core Network & Terminals**
RAN WG1 • **Radio Layer 1**	**SA WG1** • **Services**	**CT WG1** • **MM/CC/SM (Iu)**
RAN WG2 • **Radio Layer 2** • **Radio Layer 3 RR**	**SA WG2** • **Architecture**	**CT WG3** • **Interworking with external Networks**
RAN WG3 • **Iub, Iur, Iu** • **UTRAN O&M**	**SA WG3** • **Security** • **Lawful Interception**	**CT WG4** • **MAP/GTP/BCH/SS**
RAN WG4 • **Radio Performance** • **Protocol Aspects**	**SA WG4** • **Codecs**	**CT WG6** • **Smart Card Application Aspects**
RAN WG5 • **Mobile Terminal** • **Conformance Testing**	**SA WG5** • **Telecom Management**	
RAN WG6 • **Legacy RAN Radio and Protocol**	**SA WG6** • **Mission-critical Applications**	
RAN AH1 • **RAN Ad Hoc Group on ITU-R**		

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 perspec-

tive. 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 [56].

In line with this common approach to 3GPP, Release 15 (5G Phase 1) 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), now with a target date of the end 2020 due to the COVID-19 pandemic, is underway, with a 5G system based on the standards of Release 16 to meet the IMT-2020 target system defined by the ITU. The time sequences and relationships 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.

Work on Release 17 has begun. Standardization, also delayed by the COVID-19 pandemic, is expected to be completed in spring 2022 [58; 62].

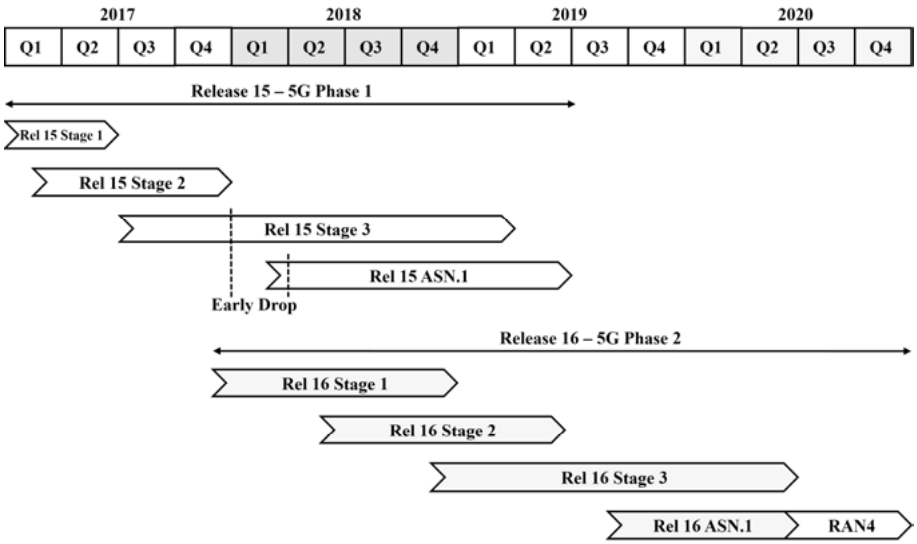

Fig. 5.2: Standardisation procedures for 3GPP for Releases 15 and 16 [57]

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 with their own, geographically limited 5G network.

Further frequency auctions for 5G by the BNetzA are planned for 2022 and 2030, as frequency ranges that become free can then be used [83].

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

After this more detailed look at 5G regulation in Germany, Switzerland, and Austria, Table 5.6 provides an overview of 5G spectrum allocation in all EU states and the UK in mid-2020 and the near future. It should be noted that many frequency auctions planned for 2020 have been postponed due to the COVID-19 pandemic [155; 197].

Tab. 5.6: 5G spectrum assignment in the EU and UK [155; 197]

Country	Frequency range	Bandwidth	Date
Austria	700 MHz band	60 MHz	September 2020
	1,5 GHz band	90 MHz	September 2020
	2,1 GHz band	120 MHz	September 2020
	3,4 – 3,8 GHz	405 MHz	March 2019
Belgium	700 MHz band		2020
	1,5 GHz band		2020
	3,6 -3,8 GHz		2021
Bulgaria	700 MHz band		2020
	3,6 GHz band		2020
Cyprus	700 MHz band		2020
	3,6 GHz band		2020
Croatia	700 MHz band		Year-end 2021
	3,6 GHz band		Year-end 2021
	26 GHz band		Year-end 2021
Czech Rep.	700 MHz band	60 MHz	Year-end 2020
	3,4 – 3,6 GHz	190 MHz	Year-end 2020
	3,6 – 3,8 GHz	200 MHz	July 2017
Denmark	700 MHz band	80 MHz	March 2019
	900 MHz band	60 MHz	March 2019
	1,5 GHz band		Year-end 2020
	2,1 GHz band		Year-end 2020
	2,3 GHz band	60 MHz	March 2019
	3,5 GHz band		Year-end 2020
	26 GHz band		Year-end 2020
Estonia	700 MHz band		Year-end 2020
	3,6 GHz band		Year-end 2020
Finland	700 MHz band	60 MHz	November 2016
	3,4 – 3,8 GHz	390 MHz	October 2018
	25,1 – 27,5 GHz	2,4 GHz	June 2020
France	700 MHz band	60 MHz	2015
	3,4 – 3,8 GHz	310 MHz	October 2020
Germany	700 MHz band	60 MHz	2015
	1,9 GHz band	60 MHz	June 2019
	2,1 GHz band	60 MHz	June 2019
	3,4 – 3,7 GHz	300 MHz	June 2019

Country	Frequency range	Bandwidth	Date
Greece	700 MHz band		Year-end 2020
	2,1 GHz band		Year-end 2020
	3,6 GHz band		Year-end 2020
	26 GHz band		Year-end 2020
Hungary	700 MHz band	50 MHz	March 2020
	3,4 – 3,8 GHz	310 MHz	March 2020
Ireland	700 MHz band		Q4 2020
	2,1 GHz band		Q4 2020
	2,3 GHz band		Q4 2020
	2,6 GHz band		Q4 2020
	3,4 – 3,8 GHz	360 MHz	May 2017
Italy	700 MHz band	75 MHz	September 2018
	3,6 – 3,8 GHz	200 MHz	October 2018
	26,5 – 27,5 GHz	1 GHz	October 2018
Latvia	3,4 – 3,7 GHz	100 MHz	November 2017
	3,55 – 3,6 GHz	50 MHz	September 2018
Lithuania	700 MHz band		Year-end 2020
	3,4 – 3,8 GHz		Year-end 2020
Luxembourg	700 MHz band	60 MHz	July 2020
	3,4 – 3,7 GHz	330 MHz	July 2020
	26 GHz band		Year-end 2020
Malta	700 MHz band		June 2021
	3,4 – 3,8 GHz		Year-end 2020
	26 GHz band		Year-end 2020
Netherlands	700 MHz band	60 MHz	Started June 2020
	1,5 GHz band	40 MHz	Started June 2020
	2,1 GHz band	120 MHz	Started June 2020
	26 GHz band		2021
Poland	3,6 – 3,8 GHz		Year-end 2020
	26 GHz band		2021
Portugal	700 MHZ band		Year-end 2020
	900 MHz band		Year-end 2020
	1,8 GHz band		Year-end 2020
	2,1 GHz band		Year-end 2020
	2,6 GHz band		Year-end 2020
	3,6 GHz band		Year-end 2020
	3,4 – 3,6 GHz	100 MHz	October 2019
Romania	700 MHz band		Year-end 2020
	800 MHz band		Year-end 2020
	1,5 GHz band		Year-end 2020
	2,6 GHz band		Year-end 2020
	3,4 – 3,8 GHz		Year-end 2020
	3,4 – 3,6 GHz	110 MHz FDD	2016
	3,6 – 3,8 GHz	145 MHz TDD	2015

Country	Frequency range	Bandwidth	Date
	26 GHz band		2021
Slovenia	700 MHz band		Q1 2021
	1,5 GHz band		Q1 2021
	2,1 GHz band		Q1 2021
	2,3 GHz band		Q1 2021
	3,5 – 3,8 GHz		Q1 2021
	26 GHz band		Q1 2021
Slovakia	700 MHz band		Postponed
	900 MHz band		Postponed
	1,8 GHz band		Postponed
	3,4 – 3,6 GHz	60 MHz	August 2015
	3,6 – 3,8 GHz	40 MHz	October 2017
Spain	700 MHz band		Q1 2021
	3,4 – 3,6 GHz	160 MHz	2016
	3,6 – 3,8 GHz	200 MHz	July 2018
Sweden	700 MHz band	60 MHz	December 2018
	2,3 GHz band		Year-end 2020
	3,4 – 3,7 GHz		Postponed
UK	700 MHz band	80 MHz	Q1 2021
	3,4 – 3,6 GHz	190 MHz	April 2018
	3,6 – 3,8 GHz	120 MHz	Q1 2021
	26 GHz band		Local licenses

According to Table 5.6, low (700 – 900 MHz), mid (1,5 – 2,6 GHz, 3,4 – 3,8 GHz), and high-frequency ranges (26 GHz band) are used across all countries mentioned. 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 refarming 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 when the 26 GHz process ended [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 5.7 [193]. It is noticeable that the focus so far has been mainly on spectra with millimeter waves, up to 48 GHz. Only recently has the FCC begun to allocate the mid 5G frequencies from 3.5 GHz upwards, which are preferred in Europe.

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 5.7) [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 2.5 GHz band, previously allocated to the Educational Broadband Service (EBS). The order eliminated restrictions on what types of entities can hold licenses in the band and created an opportunity for rural Tribal Nations to obtain licenses for unassigned 2.5 GHz spectrum covering Tribal lands. The FCC will start the planning process to auction the 2.5 GHz spectrum that remains unlicensed after the allocation to Tribal Nations is completed in August 2020. The new licenses will be 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. The auction for remaining unlicensed areas is not currently scheduled, but the auction is expected to begin in 2021 [200].

Tab. 5.7: 5G frequency auctions in the USA [193]

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

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].

To promote the introduction and rollout of 5G networks in the USA, the FCC has launched a so-called 5G FAST plan (Facilitate America's Superiority in 5G Technology) [195]. This strategy includes three key components:

- Pushing more spectrum into the marketplace
- Updating infrastructure policy
- Modernizing outdated regulations.

The FCC is taking action to make additional spectrum available for 5G services.

- High-band: The FCC has made auctioning high-band spectrum a priority. The FCC concluded its first 5G spectrum auctions in the 28 GHz band (see auction 101 in Table 5.7); the 24 GHz band (auction 102); and the upper 37 GHz, 39 GHz, and 47 GHz bands (auction 103). With these auctions, the FCC is releasing almost 5 GHz of 5G spectrum into the market – more than all other flexible use bands combined.
- Mid-band: Mid-band spectrum has become a target for 5G buildout, given its balanced coverage and capacity characteristics. With the 2.5 GHz, 3.5 GHz (auction 105), and 3.7-4.2 GHz (auction 107) bands, the FCC will make more than 600 MHz available for 5G deployments.
- Low-band: The FCC is acting to improve the use of low-band spectrum (useful for wider coverage) for 5G services, with targeted changes to the 600 MHz, 800 MHz, and 900 MHz bands.
- Unlicensed: Recognizing that the unlicensed spectrum will be important for 5G, the agency is creating new opportunities for the next generation of WLAN in the 6 GHz and above 95 GHz band.

The FCC is updating infrastructure policy and encouraging the private sector to invest in 5G networks. For this, FCC adopted new rules that will reduce federal regulatory impediments for deploying infrastructure needed for 5G. In addition, the rules for state and local review of small cells were reformed for speeding up. Besides, the FCC is modernizing outdated regulations to promote the wired backbone (e.g., with fiber networks) of 5G systems and digital opportunities for the customers [195].

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 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-

https://doi.org/10.1515/9783110724509-006

motive, or health industry. Let us spun 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 be 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 – NR, for FR1 and/or FR2 – gNB can be divided into gNB-CU and gNB-DU – Split CU into CU-UP and CU-CP – Dual Connectivity – Coexistence with LTE
5G core network – 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 – 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 – 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
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.

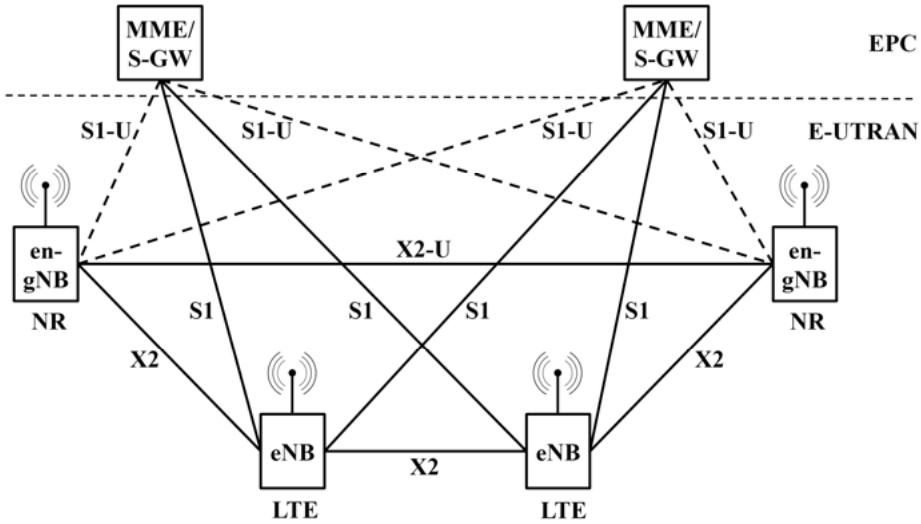

Fig. 6.1: 5G-NSA architecture [19]

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, no 4G infrastructure is necessary [19].

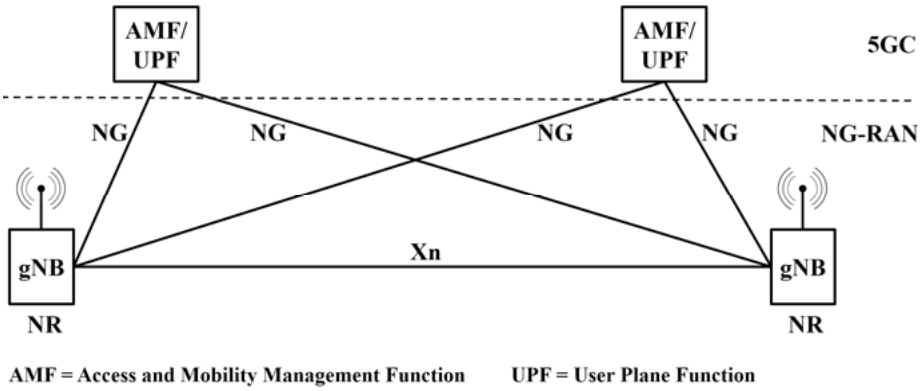

Fig. 6.2: 5G-SA architecture [19]

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.

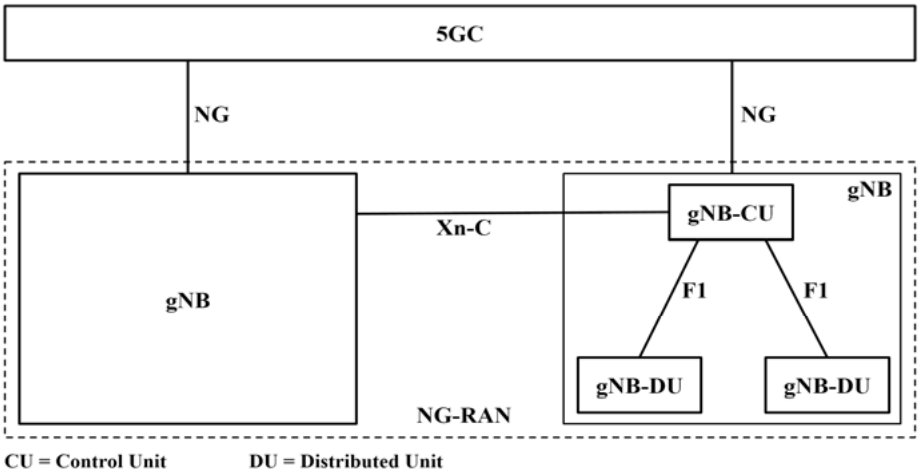

Fig. 6.3: NG-RAN architecture with split gNB [19]

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.

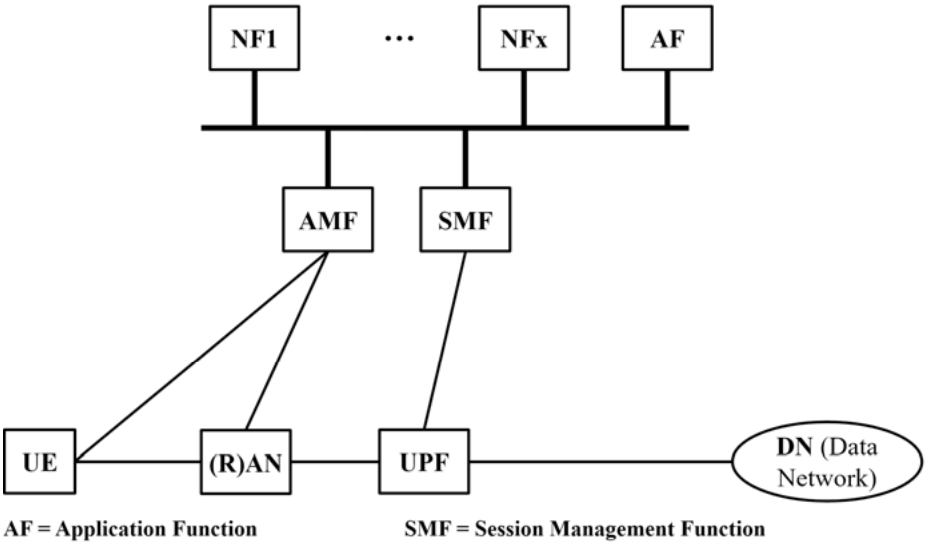

Fig. 6.4: 5G system architecture with SBA [19]

According to Table 6.1, a 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].

According to Figure 5.2, Release 16 is currently in the standardization process. As a consequence, a complete description of the features is not yet available. Table 6.2 contains a preliminary overview based on [20; 60].

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

Features
New
LAN support
TSN (Time-Sensitive Networking) with highly accurate time synchronization
V2X communication with NR side link for direct UE-UE communication as well as multicast and broadcast communication
ATSSS (Access Traffic Steering, Switch and Splitting)
Maritime communication services
NR with unlicensed frequencies
SON (Self-Organising Networks) and MDT (Minimization of Drive Tests) support for NR
Enhancements
URLLC support in 5GC and physical layer
NR for Industrial IoT
Cellular IoT (NB-IoT, eMTC)
Advanced V2X
Wireless and wireline convergence
Mobile communication system for railways 2
Public warning system
Conversational services, streaming, and TV
5G location and positioning services
Network slicing

Features
SBA (Service Based Architecture)
Reduced power consumption for UEs
Dual Connectivity (DC) with lower activation times and higher data rates
Carrier Aggregation (CA)
Support of NR DL 256 QAM for frequency range 2 (FR2)
LTE for 5G
Telecom management

Release 16 standardizes a 3GPP 5G system phase 2 with its advanced functions. This, in turn, should cover the requirements of the ITU IMT-2020 target system.

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.

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].

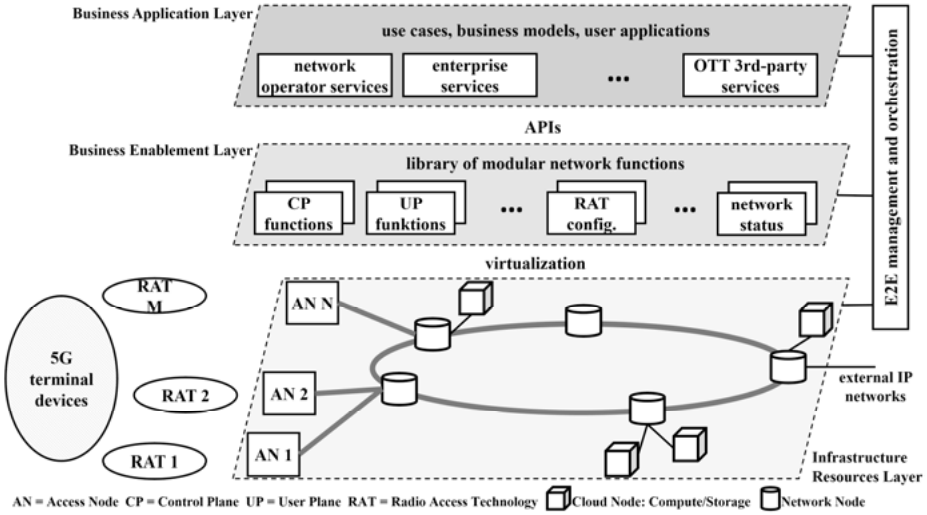

Fig. 6.5: 5G network architecture according to NGMN Alliance [96]

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

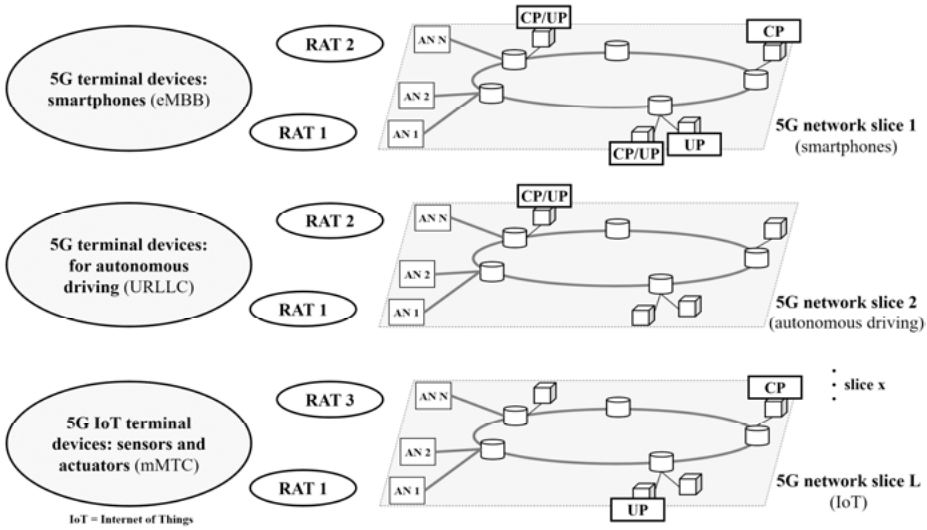

Fig. 6.6: Different 5G network slices based on the same physical infrastructure [96]

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 6.2) 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.

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 is still used in Release 15). 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 700 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.

https://doi.org/10.1515/9783110724509-007

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].

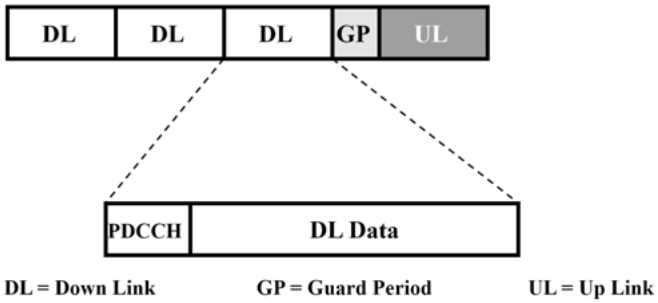

Fig. 7.1: Example of a TDD signal structure

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.

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 procedure (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.

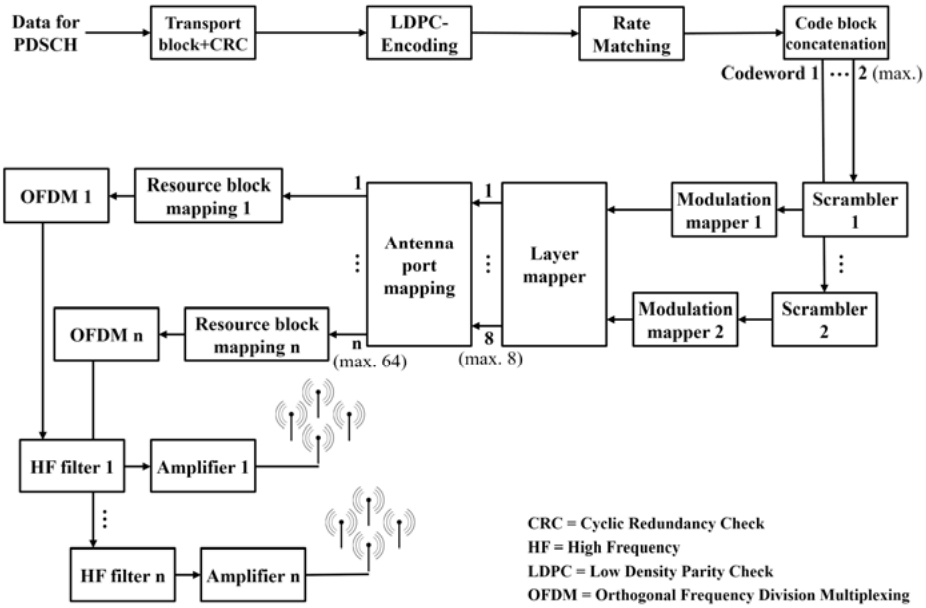

Fig. 7.2: PDSCH transmitter [106]

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 → 45^0, 01 → 315^0, 10 → 135^0, 11 → 225^0. 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].

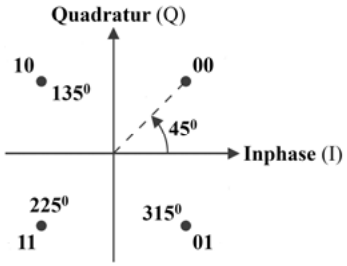

Fig. 7.3: QPSK [91]

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].

The modulation process for an NR transmitter will be explained below using the comparatively simple quadrivalent QPSK procedure. According to Figure 7.3, 2 bits each are mapped to a symbol, here a phase shift: 00 → 45^0, 01 → 315^0, 10 → 135^0, 11 →

225^0. 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 $s_{If}(t) = S \cos(2\Pi f_{If} t + \phi(t))$ 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 $s_{Hf}(t) = S \cos(2\Pi(f_{If} + f_{Hf})t + \phi(t))$ with the desired phase changes representing the bit combinations to be sent. As will be shown below, this two-step approach is very advantageous for the realization of OFDM [91].

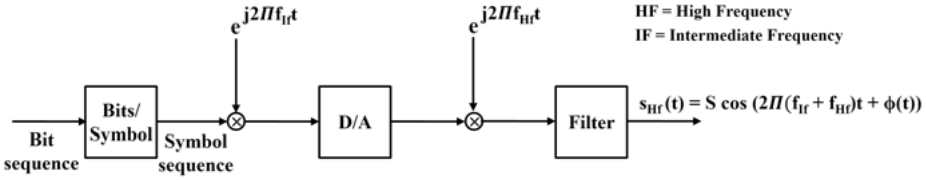

Fig. 7.4: QPSK modulation

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. A 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.

An advantage of OFDM is not only that the requirements on the transmission channel per subcarrier are significantly reduced due to the comparatively low sym-

bol rate. Additionally, particularly disturbed frequency ranges with still free subcarriers can simply be omitted, i.e., the associated subcarriers are not used.

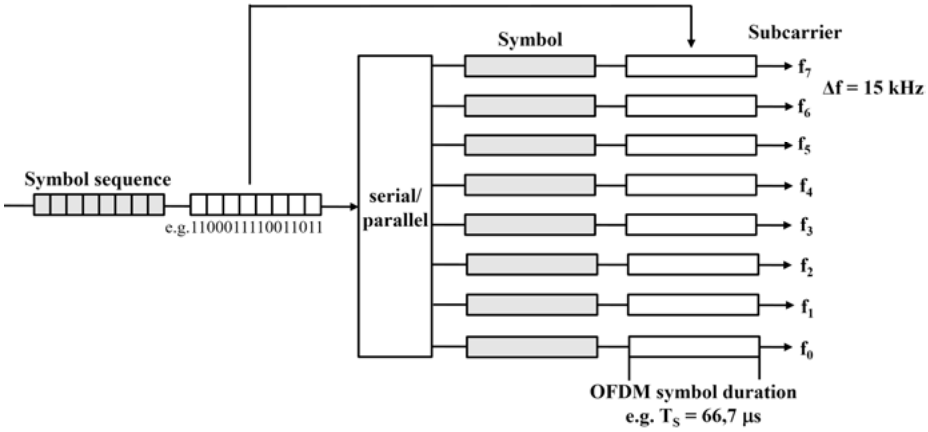

Fig. 7.5: OFDM basic principle [91]

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.

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].

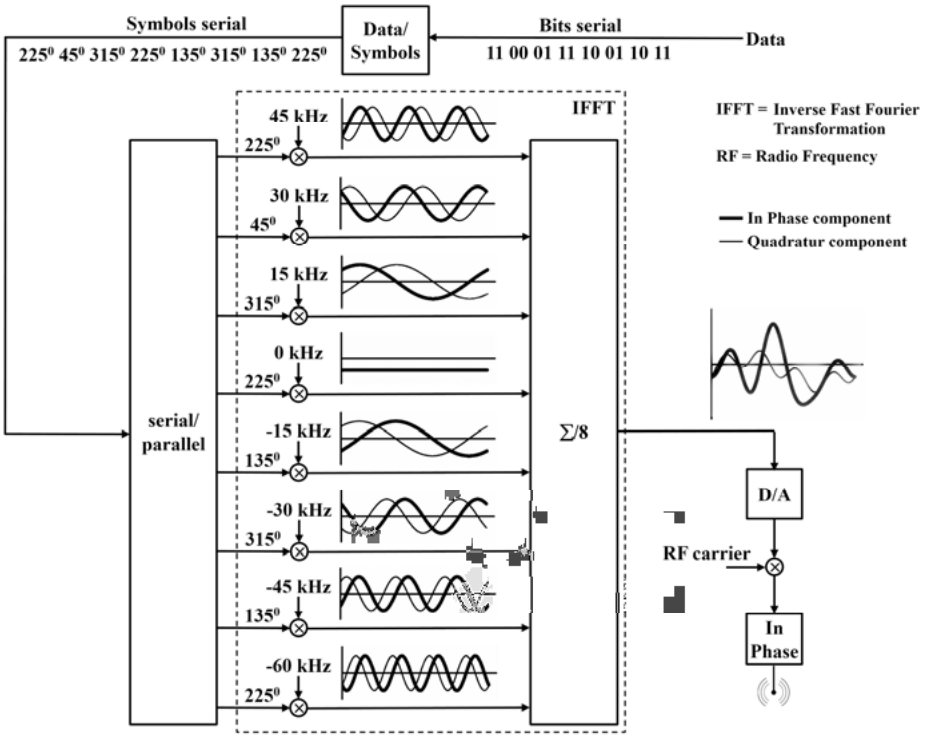

Fig. 7.6: OFDM processing [91]

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 µs, a total of approx. 71.4 µs 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].

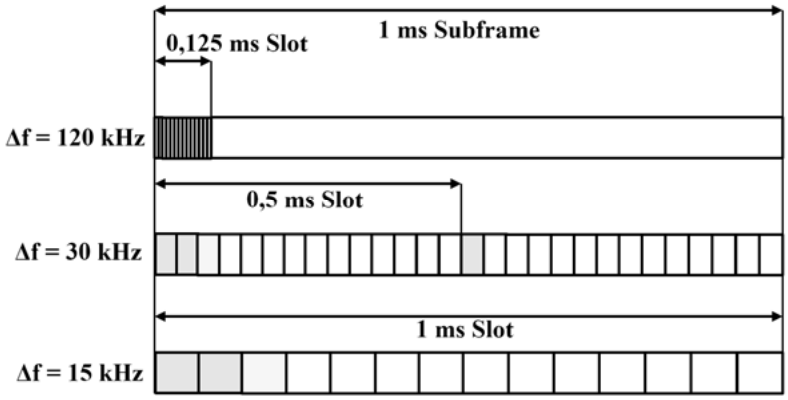

Fig. 7.7: Frame with OFDM symbols [106]

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 to 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 $24 \cdot 12 \cdot 15$ kHz = 4,32 MHz. With 275 RBs and 60 kHz subcarriers, the result is $275 \cdot 12 \cdot 60$ kHz = 198 MHz. Conversely, for a bandwidth of approx. 20 MHz, using 30 kHz subcarriers, about 56 RBs are necessary according to 20 MHz/(12 · 30 kHz).

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

Δf [kHz]	OFDM symbols/ slot	Slot/ subfra- me	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.

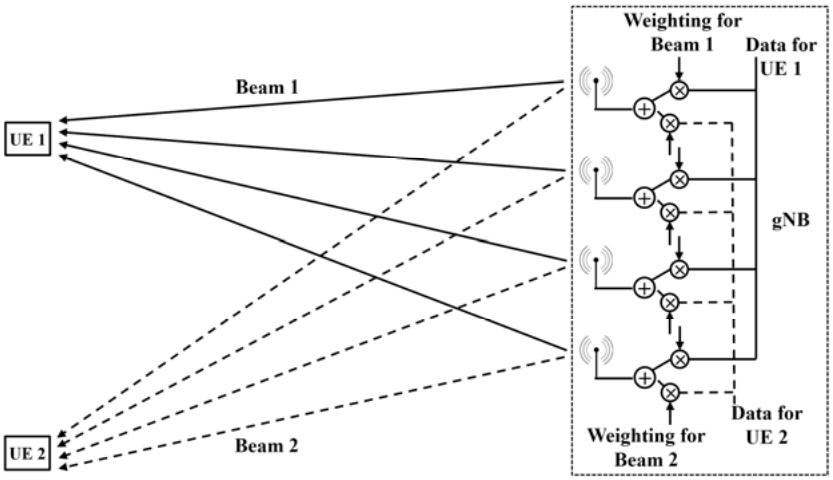

Fig. 7.8: Beamforming [91]

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].

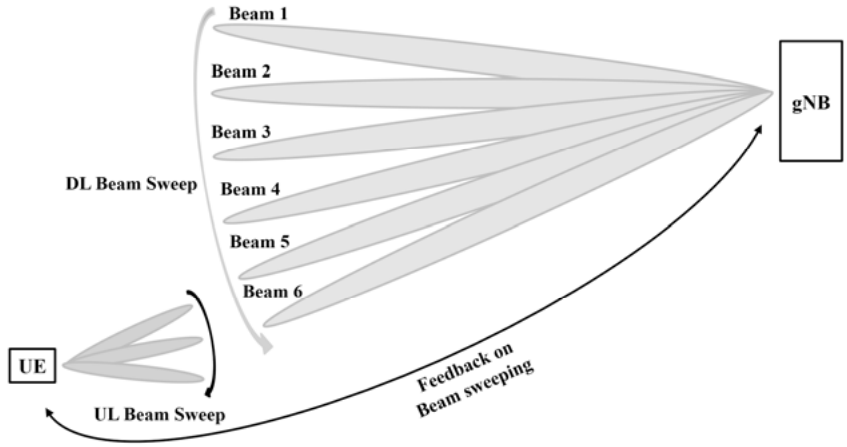

Fig. 7.9: Beam Sweeping [106]

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.

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 only associated with comparatively much lower costs [40; 169].

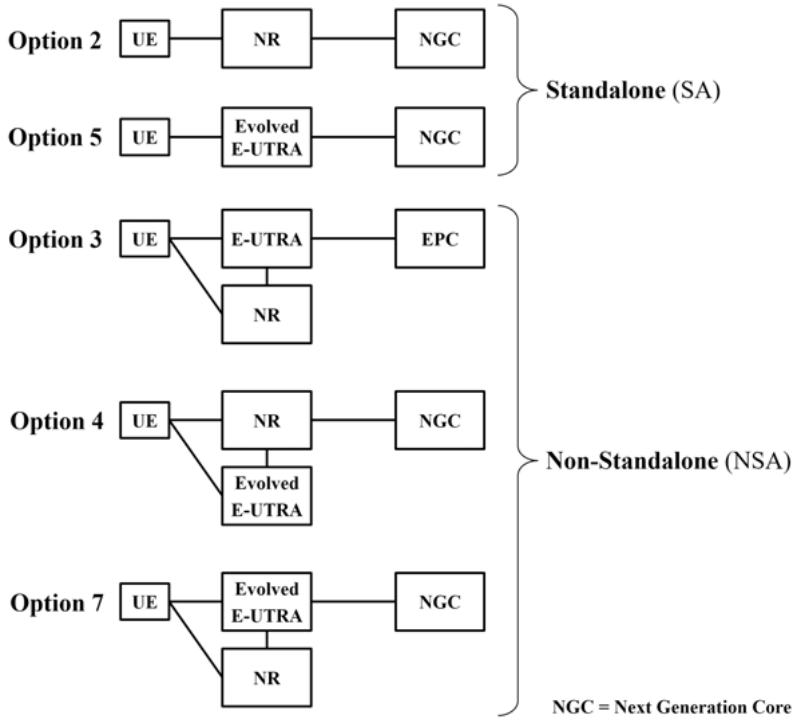

Fig. 7.10: RAN options [40]

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 anten-

na 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].

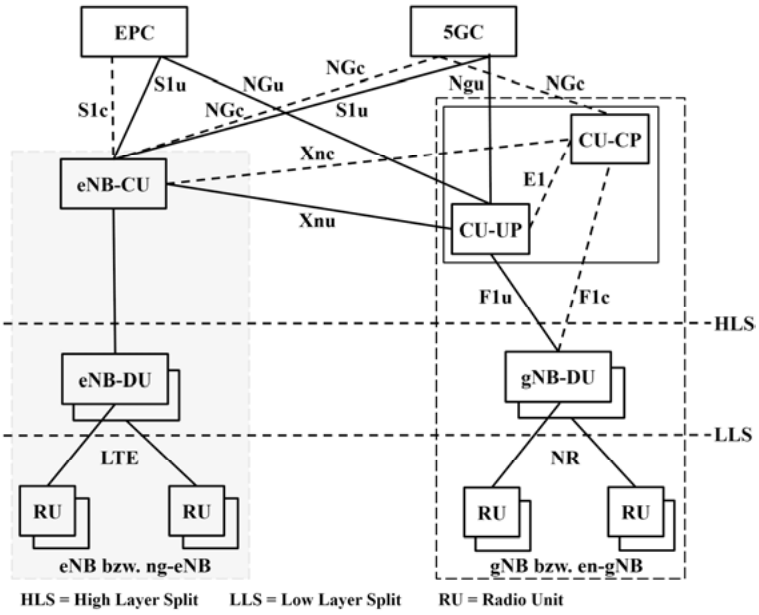

Fig. 7.11: RAN architecture for 5G, 4G and 5G/4G [129; 19]

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 easily 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 of 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), 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.

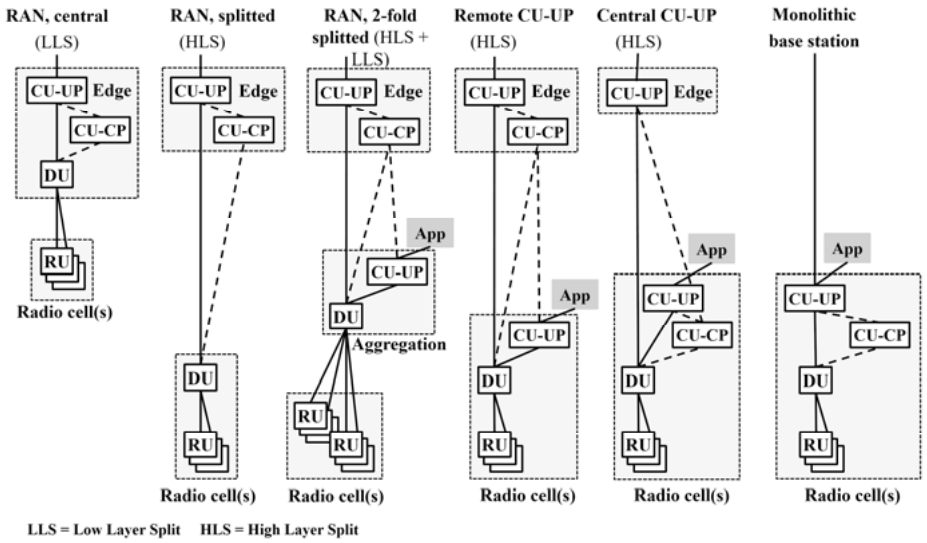

Fig. 7.12: Possible 5G-RAN architectures with function splitting [129]

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].

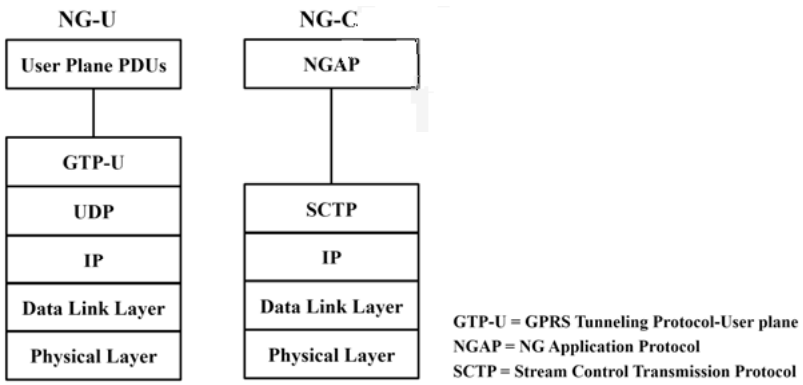

Fig. 7.13: NG protocol stacks between NG-RAN and 5G core [51]

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].

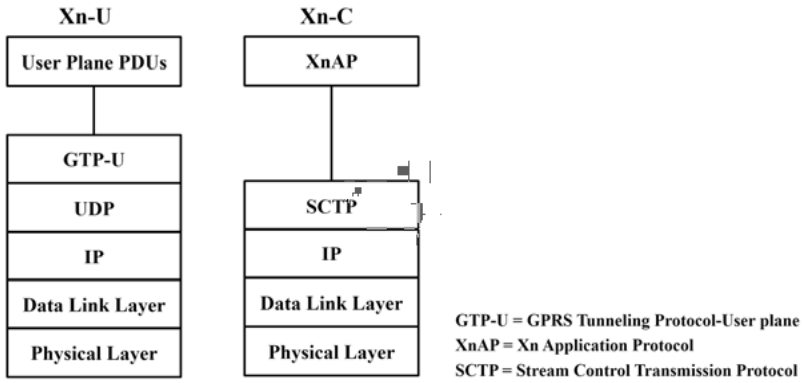

Fig. 7.14: Xn protocol stacks between gNBs in NG-RAN [51]

Besides, Figure 7.15 shows the protocol stacks for communication with a UE for both the user and the control plane. In both cases, the lower layers are realized by PDCP (Packet Data Convergence Protocol), RLC (Radio Link Control), and MAC (Medium Access Control). The RLC protocol ensures reliable communication on the air interface in layer 2. Finally, the PDCP is primarily responsible for transporting user and signaling data, including encryption and integrity assurance.

In the case of the UP protocol stack, the SDAP (Service Data Adaptation Protocol) assigns a corresponding QoS flow and a corresponding Data Radio Bearer (DRB) to the transmitted user data packets. In the CP, the RRC protocol (Radio Resource Control) is used to establish and terminate signaling connections between the gNB and UE, provides Signaling Radio Bearer (SRB), paging, mobility control in the event of handover and QoS management. As shown in Figure 7.15, NAS (Non Access Stratum) messages are exchanged between a UE and the 5GC function AMF (Access and Mobility Management Function) based on an RRC connection to establish and maintain sessions [51].

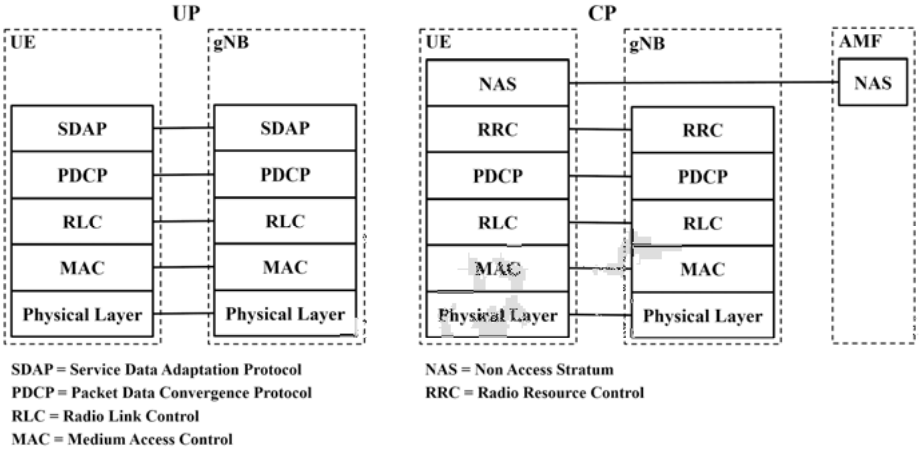

Fig. 7.15: UP and CP protocol stacks between UE and NG-RAN or 5GC [51]

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].

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 6.2 and applied to the RAN above, the O-RAN Alliance aims to create an open and intelligent 5G RAN as a basis for smaller manufacturers and operators' developments. 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 7.16 shows the O-RAN reference architecture with the interfaces and functionalities from Figures 7.11 and 7.13 to 7.15. Also, there are other interfaces to be specified by the O-RAN Alliance and, extending the separation of UP and CP, the introduction of a RIC that can be orchestrated and managed via an NFV MANO system (see Section 3.1) [109].

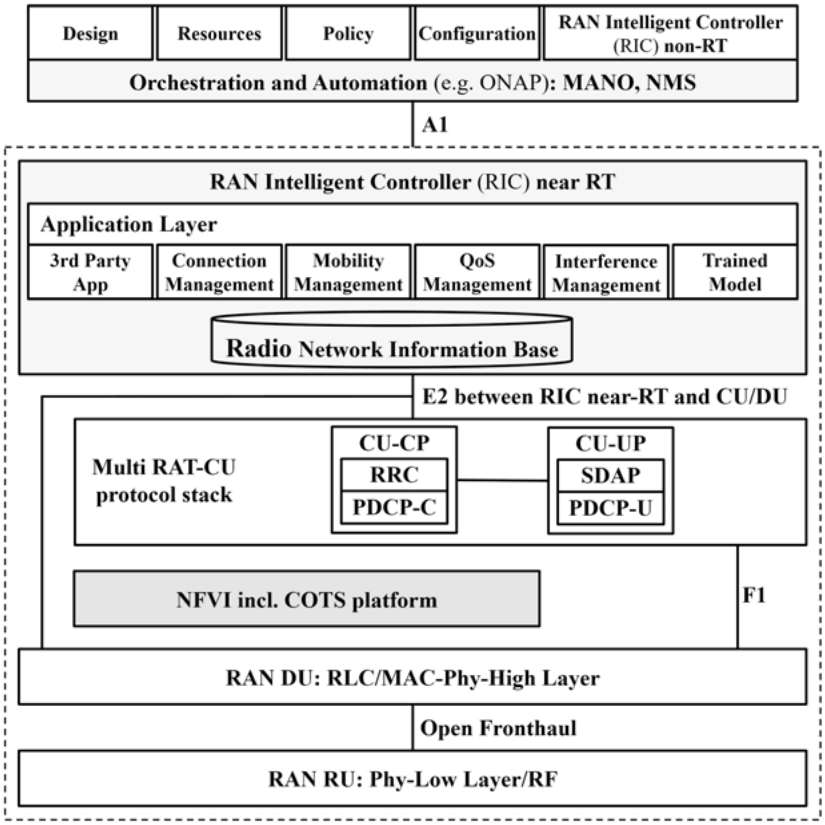

Fig. 7.16: O-RAN reference architecture [109]

Politics in the USA promote such an open RAN solution. The main reason given is the improved IT security provided by Open-RANs. A corresponding US law would make it considerably more difficult for the leading manufacturers in the world market for 5G systems with their previously proprietary RAN systems to survive in the US market. Instead, it would enable Open-RAN vendors to enter the 5G market on a larger scale.

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.

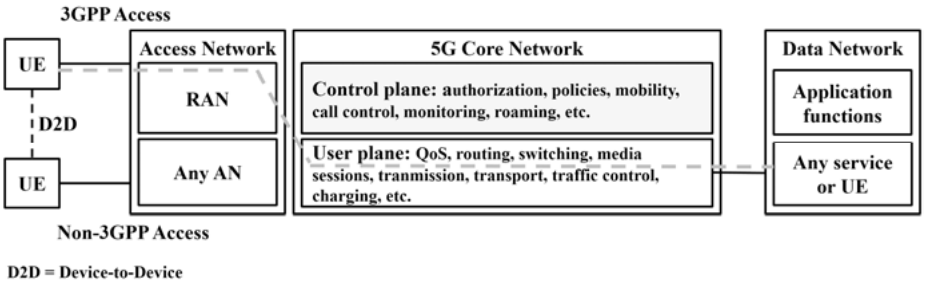

Fig. 8.1: Structuring of a 5G system

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, provision 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

https://doi.org/10.1515/9783110724509-008

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. Section 8.1 gives 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 8.2 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 8.3.

8.1 Core Network Functions

The 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.

Consequently, the following network functions were standardized in Release 15 according to [37]: Authentication Server Function (AUSF), Access and Mobility Management Function (AMF), Unstructured Data Storage Function (UDSF), Network Exposure Function (NEF), Network Repository Function (NRF), Network Slice Selection Function (NSSF), Policy Control Function (PCF), Session Management Function (SMF), Unified Data Management (UDM), Unified Data Repository (UDR), User Plane Function (UPF), Application Function (AF), 5G-Equipment Identity Register (5G-EIR), Security Edge Protection Proxy (SEPP), Network Data Analytics Function (NWDAF), Charging Function (CHF). Besides, there are the connected networks – (Radio) Access Network ((R)AN) and Data Network (DN) for services of the network operator, Internet access, or for 3rd party services – and the User Equipment (UE).

Figure 8.2 shows some of the listed network functions in a representation with reference points Nx, i.e., these NFs communicate peer-to-peer via standardized reference points. The UPF (User Plane Function) is responsible for user data handling, i.e., the PDU sessions (Protocol Data Unit), in 5GC. It represents the anchor point for traffic via N3 (NG-U) from and to AN for mobile use and the transition to DN. Their tasks include routing and forwarding of data packets, including traffic control and redirection, QoS handling (including ensuring DL and UL bit rates, marking packets, assigning packets to flows), packet inspection, lawful interception (LI), and the collection and provision of usage data. The UPF functionality can be programmed by the SMF via reference point N4.

At the N1 reference point, the AMF (Access and Mobility Management Function) in the CP terminates the NAS session signaling with the UE and at the N2 reference point the signaling (NG-C) with the (R)AN (see Section 7.2). Registration, connection, accessibility, and mobility management, including authentication and authorization, are also carried out here. Besides, lawful interception and localization data are recorded here.

The SMF (Session Management Function) is mainly responsible for session control, i.e., establishing, 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. Also, data for charging and also for LI are captured here.

The PCF (Policy Control Function) provides uniform policies for network behavior (e.g., on QoS, traffic forwarding, AN priorities) and makes these available to other NFs. To this purpose, it accesses the user profiles and application requirements in the UDF (Unified Data Repository).

The UDM (Unified Data Management) handles authentication and identification of data based on user profiles. The AUSF (Authentication Server Function) represents an authentication server, and the NSSF (Network Slice Selection Function) selects the network slice(s) responsible for a UE and determines the AMF instance(s) for it. The AF (Application Function) finally represents an additional CP function,

which is, e.g., provided by a 3rd party user. It is not part of the 5GC but can communicate with the NFs of the 5GC via a NEF (Network Exposure Function) [37].

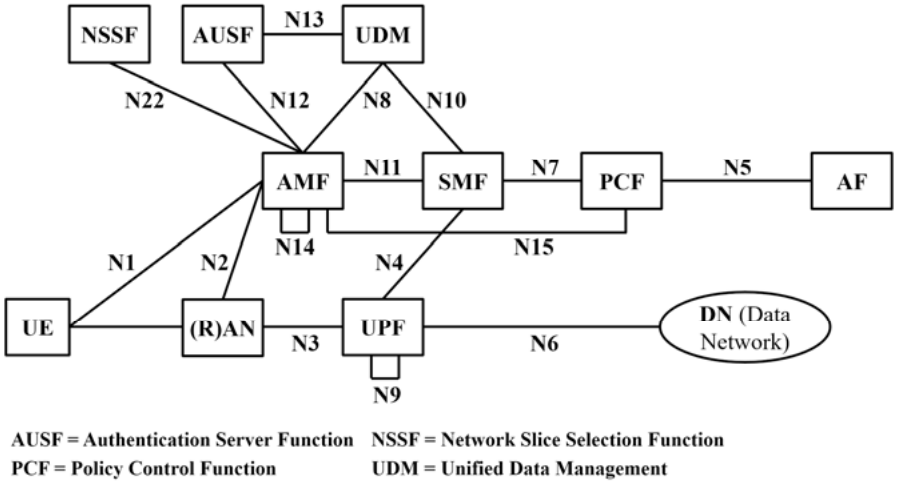

Fig. 8.2: Network functions with reference points for a 5G system without roaming [37]

As a comparison of the above list with the NFs shown in Figure 8.2 shows, there are additional network functions that have not yet been explained. These interact with the other NFs not via specially specified reference points, but via an API generally standardized for NF communication in the 5GC as the backbone of the SBA (see Section 8.2). The network functions NRF and NEF are also crucial for further understanding.

With the NRF (Network Repository Function), the available NFs are known with their network services and can be queried by other NFs who want to use their services (Discovery). Finally, the NEF (Network Exposure Function) collects, stores, and provides the available services with its functions for NFs and 3rd party users.

Before we discuss the interaction of the described essential network functions in more detail, the SBA concept will be explained in Section 8.2. But first, Figure 8.3 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.

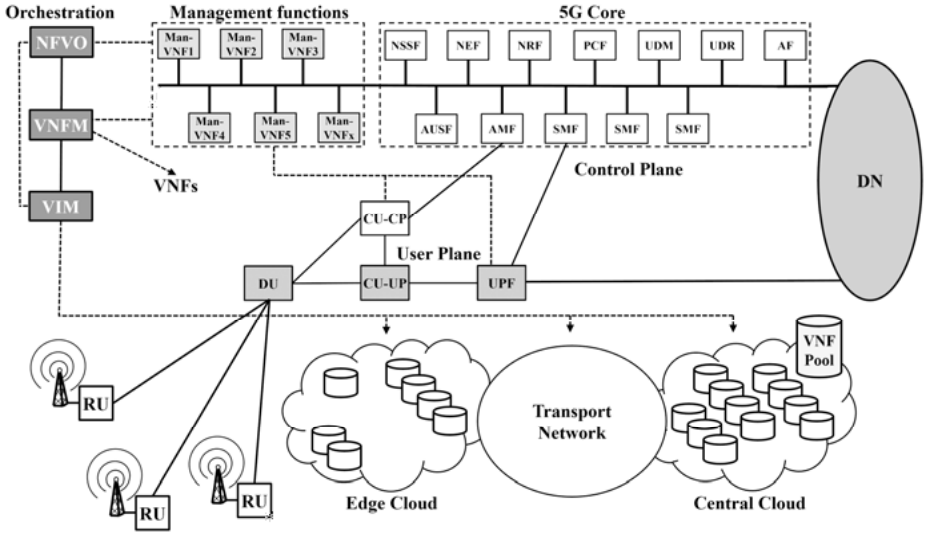

Fig. 8.3: Network functions for the user plane, control plan, network management, and orchestration in a 5G system

8.2 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 8.4 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 8.2. 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), Nnssf (NSSF), Nausf (AUSF), Nudr (UDR), Nudsf (UDSF), N5g-eir (5G-EIR), Nnwdaf (NWDAF), Nchf (CHF).

For completeness, the reference points from Figures 8.2 and 8.4 are also given. 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), N5 (PCF – AF), N6 (UPF – DN), N7 (SNF – PCF), N8 (UDM – ADF), N9 (UPF – UPF), N10 (UDM – SMF), N11 (AMF – SMF), N12 (AMF – AUSF), N13 (UDM – AUSF), N14 (AMF – AMF), N15 (PCF – AMF), N22 (AMF – NSSF) [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].

An example will illustrate the interaction of selected NFs in Figure 8.4. 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. Figure 8.5 shows the sequence for this.

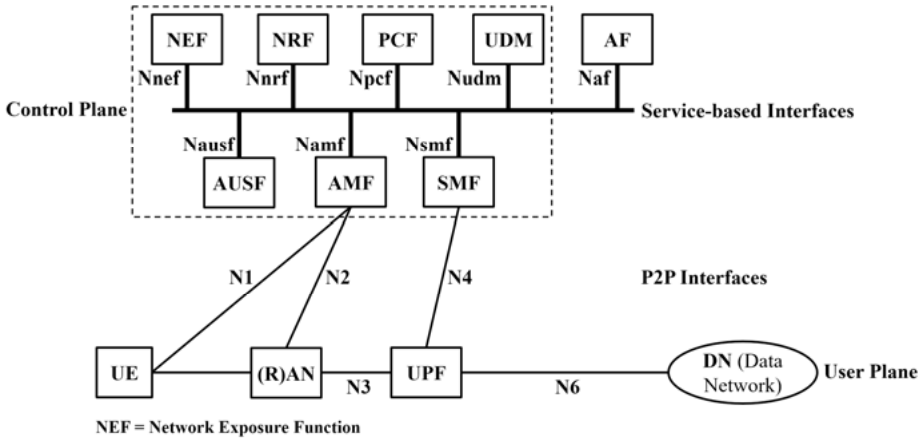

Fig. 8.4: Network functions with SBA interfaces for a 5G system without roaming [37]

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 8.4 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].

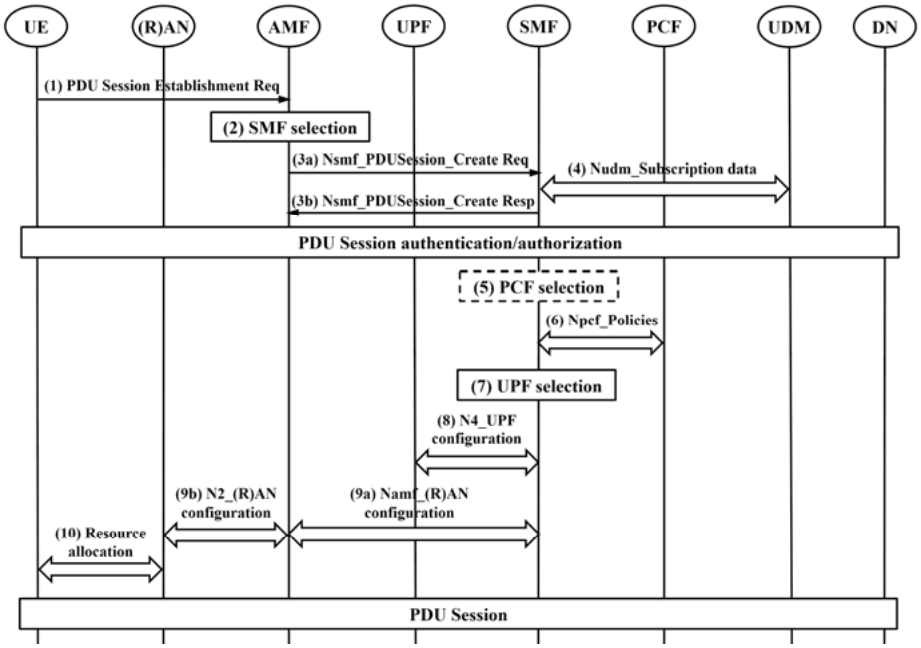

Fig. 8.5: PDU session setup initiated by UE [39]

According to Figure 8.5, 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, user data packets can be transmitted from the UE to the DN (Data Network) and vice versa [39; 88].

The advantages of the SBA approach, with its modularity, openness, and flexibility, also apply to roaming. Figure 8.6 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, how-

ever, 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].

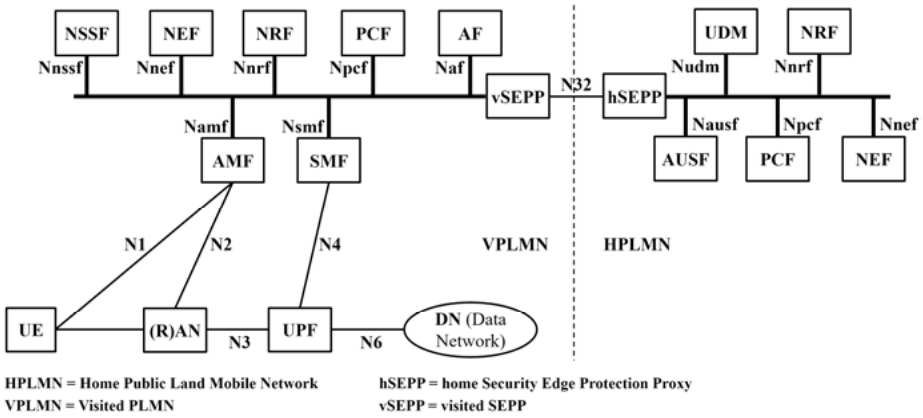

Fig. 8.6: Roaming with local termination on the visited network [37]

Another advantage of the SBA solution chosen for the 5G core is the separation of processing (computing) from data storage. As shown in Figure 8.7, 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].

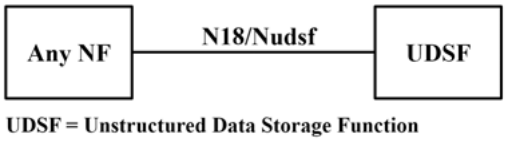

Fig. 8.7: Data storage with UDSF (Unstructured Data Storage Function) [37]

Also, the 5G core provides specific data storage for certain NFs with the UDR (Unified Data Repository), as shown in Figure 8.8. 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].

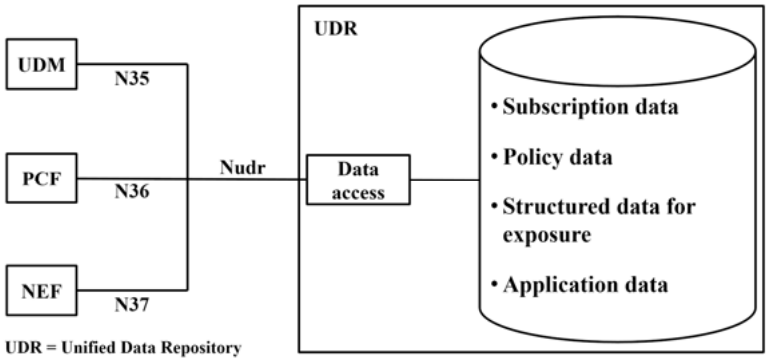

Fig. 8.8: 5GC data storage architecture for SBA [37]

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.
- 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 here 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 8.9, 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].

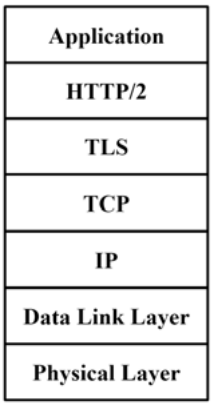

Fig. 8.9: SBI protocol stack [43]

As already mentioned in Section 8.1 and illustrated in Figure 8.4, 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. However, this is only possible indirectly via a NEF (Network Exposure Function). Advantageously, a RESTful API is also used here – as shown in Figure 8.10 [135].

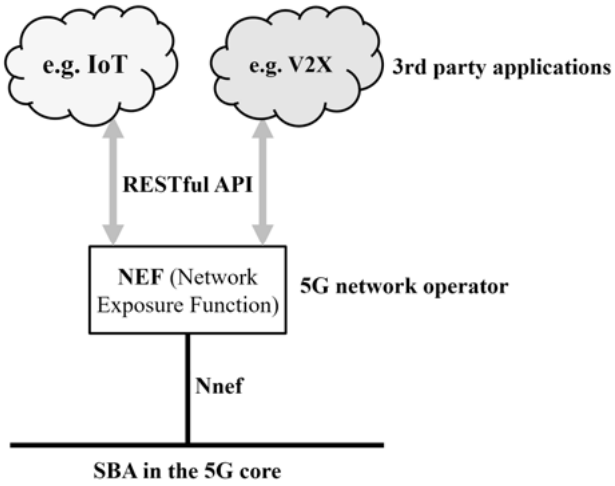

Fig. 8.10: RESTful API for connecting applications to the 5GC via a NEF [135]

With the help of Figure 8.5, 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 8.11 summarizes three example scenarios [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 does so by sending 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 8.5, 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.

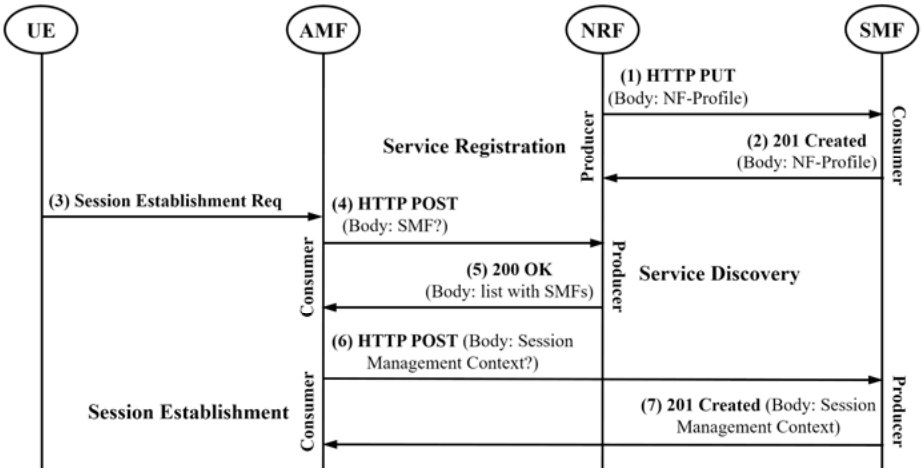

Fig. 8.11: Examples for the interaction of NFs via RESTful APIs in the SBA [135; 39]

8.3 Network Slicing

A fundamental 5G design principle (see Section 6.1) is the Service Based Architecture described in Section 8.2. 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, in order to operate several, here 5, logical communication networks with different characteristics in parallel on one physical network platform.

Figure 8.12 illustrates dependencies. The physical network consists of the 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 virtual-

ization 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 of 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.

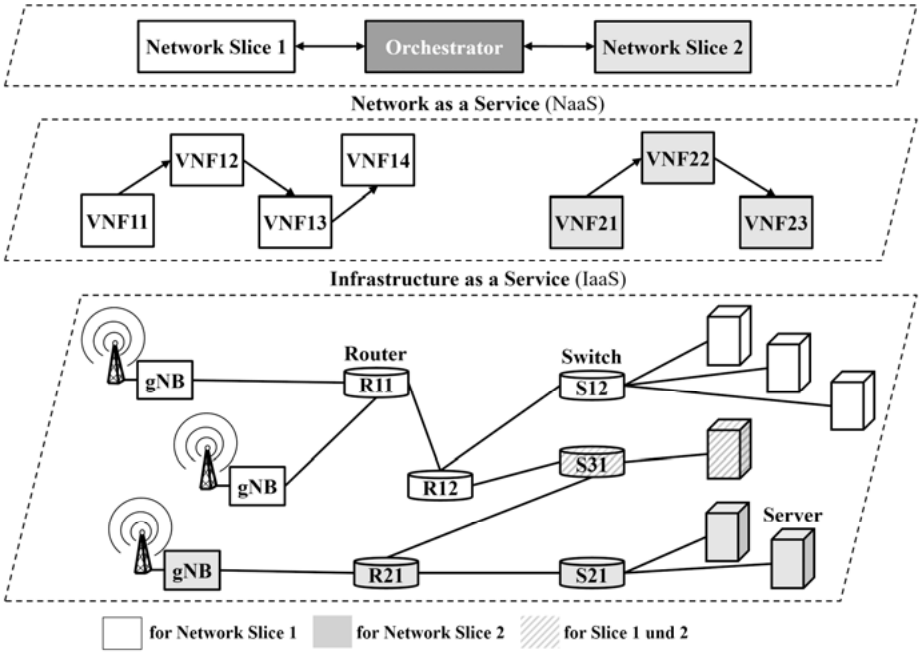

Fig. 8.12: Network slicing based on common physical network infrastructure and virtualization [133]

Figure 8.13 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).

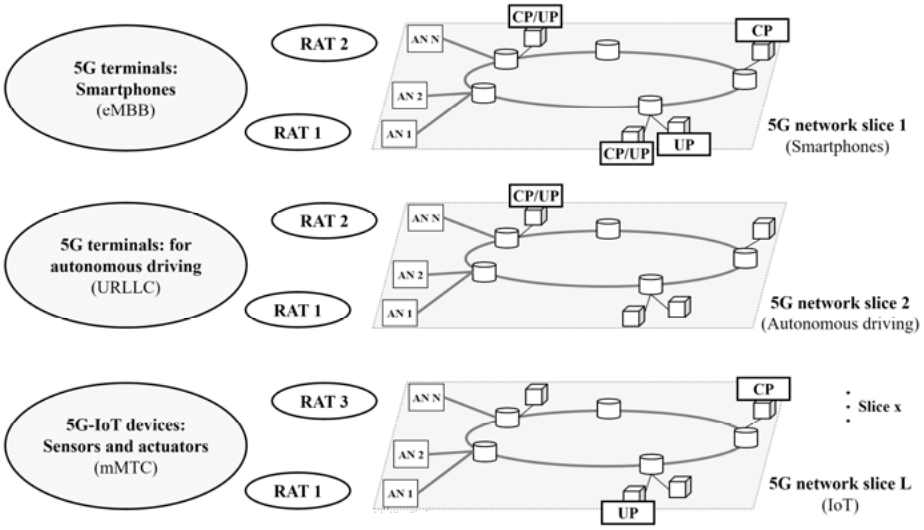

Fig. 8.13: Application scenarios for network slices [140]

Figure 8.14 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.

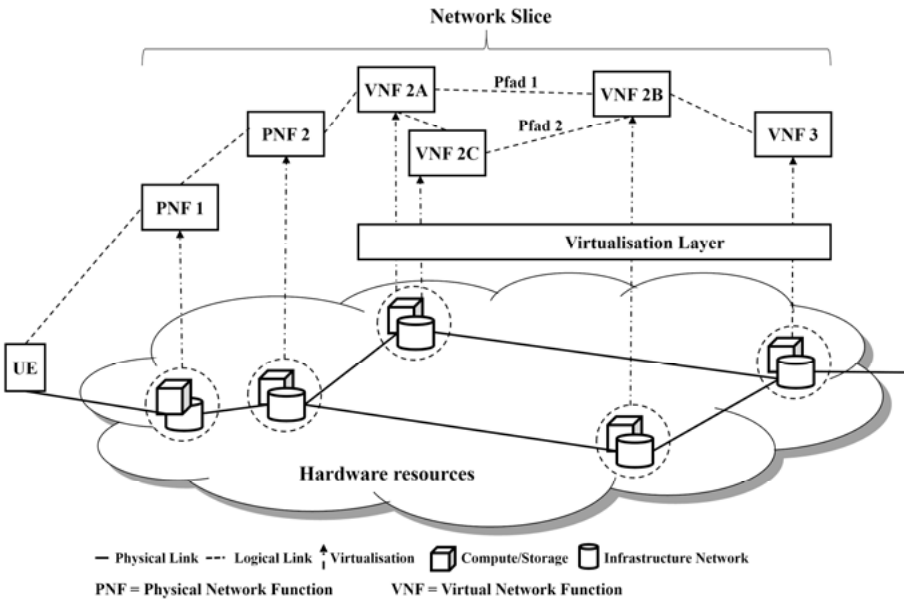

Fig. 8.14: Network slice with orchestrated VNFs and PNFs [140]

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, for 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.

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, first an AMF and then the combination(s) of SMF and UPF must be selected (see Section 8.2).

Based on these introductory notes, a network with three exemplary network slices, as shown in Figure 8.15, 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 NSSAI (Network Slice Selection Assistance Information) and DNN (Data Network Name). The latter identifies the target DN [103].

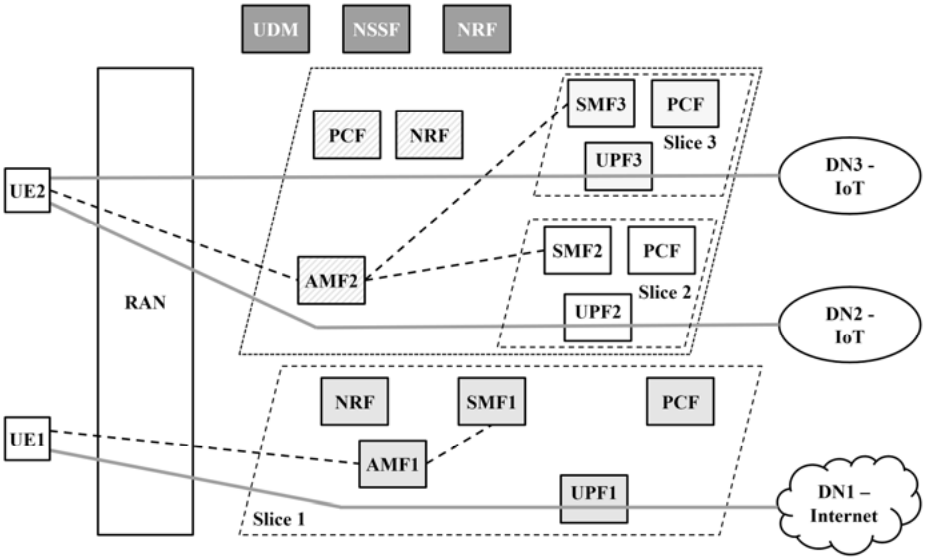

Fig. 8.15: Examples for network slicing in 5G core [103]

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 8.1 and 8.2, the NRF (Network Repository Function) is used to easily find a required NF, e.g., SMF, UPF, and PCF. As shown in Figure 8.15, 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 8.15 does not show the RAN slicing supported in principle by the 5GC. 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 sys-

tem, 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 8.16 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 network 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 will be changed. In (8), the NSSF then 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].

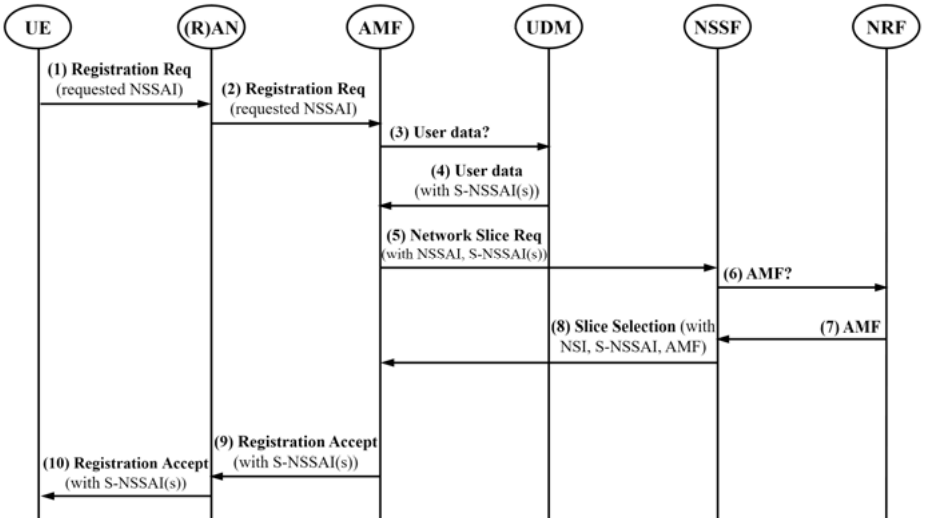

Fig. 8.16: Selecting a network slice [88; 39]

Finally, Figure 8.17 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. Togeth-

er, 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].

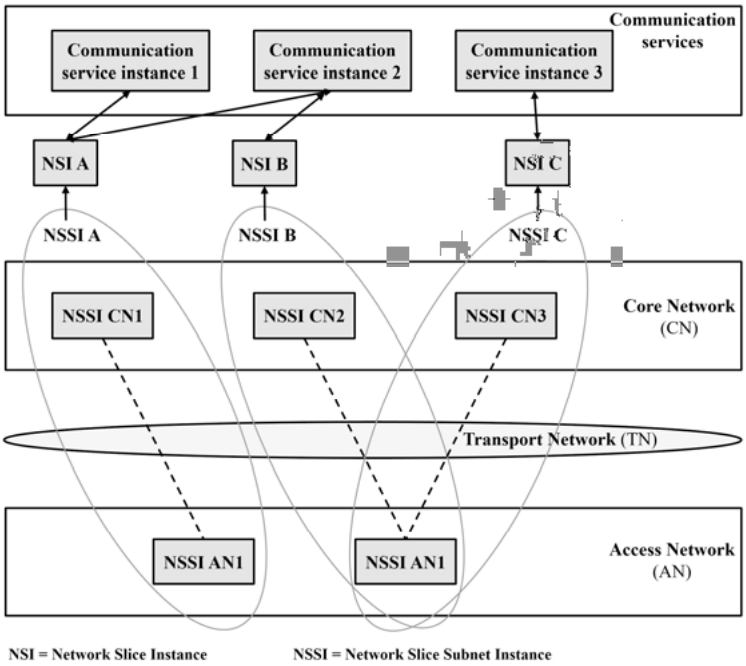

Fig. 8.17: Network Slices formed from Network Slice Subnet Instances (NSSI) [42]

9 5G System

Concerning 5G, RAN has been discussed in more detail in Section 7.2 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 will also become clear that the 5th generation will finally leave the field of mobile communications networks and become 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 then 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-

https://doi.org/10.1515/9783110724509-009

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, will be phased out first.

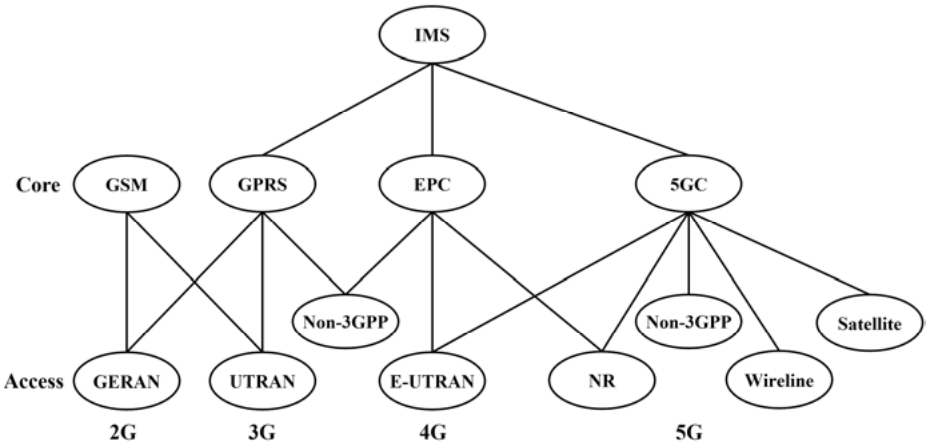

Fig. 9.1: Migration of mobile networks towards 5G

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). Finally, this chapter will conclude with an overall assessment of a 5G system in a comprehensive view of the various concepts and technologies combined in it (see Section 9.4).

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 7.11, 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 NGu 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 [129; 19].

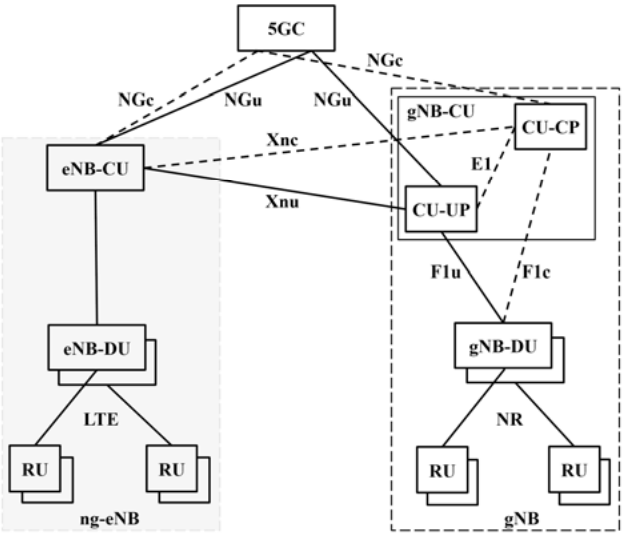

Fig. 9.2: NG-RAN with NR and LTE [129]

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) and PGW (Packet Data Network Gateway) 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].

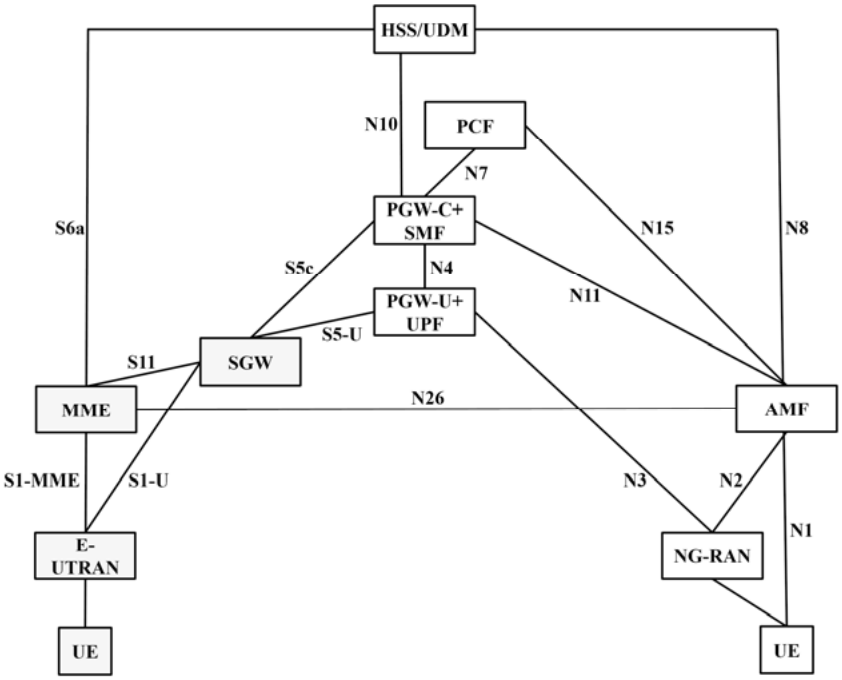

Fig. 9.3: 4G/5G migration through interworking between EPC and 5G core [37]

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 stand-alone system. A network operator could then have a parallel 4G and 5G solution, which would then be merged using Options 4 or 7.

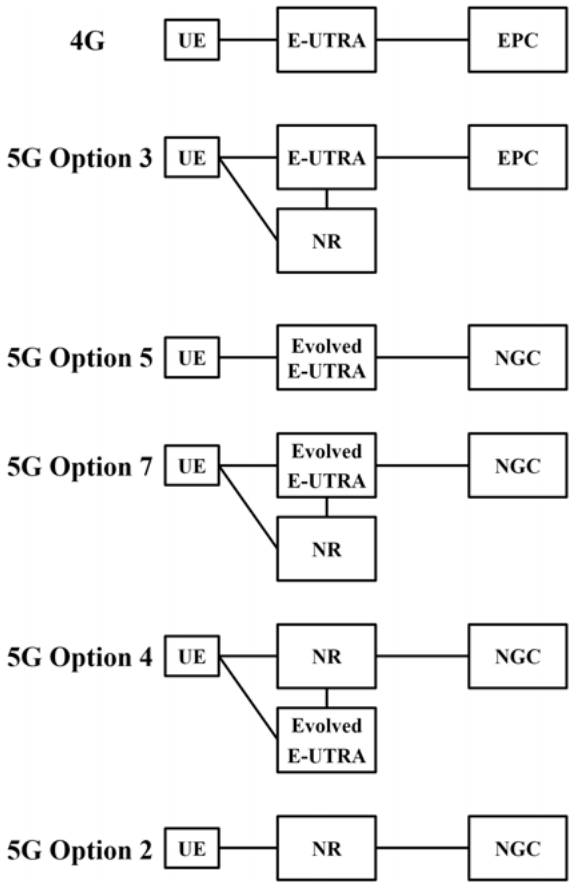

Fig. 9.4: RAN migration from 4G to 5G

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. 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].

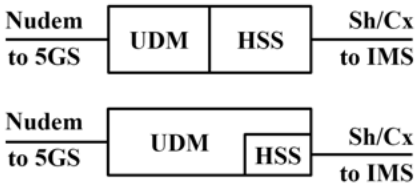

Fig. 9.5: Combination of HSS and UDM for IMS integration [33]

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 or a gateway) [33].

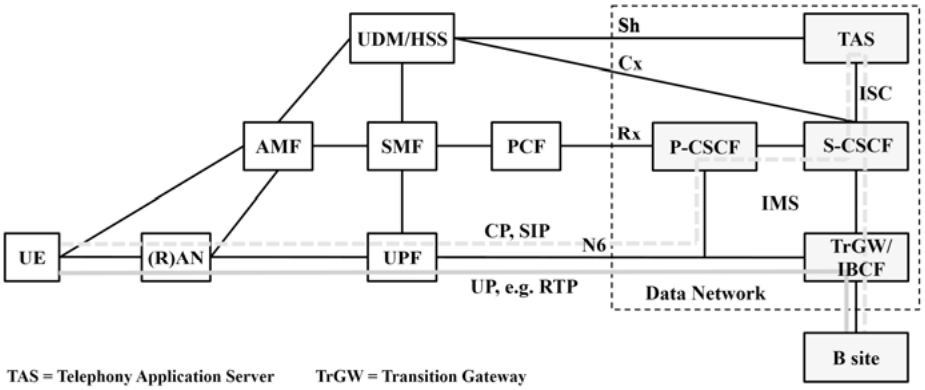

Fig. 9.6: Multimedia over IP with IMS in 5G system [33]

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 removes 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 7.2, 8.1 and 8.2). As a result, AN and CN harmonize directly here; additional functions for interconnection are not necessary.

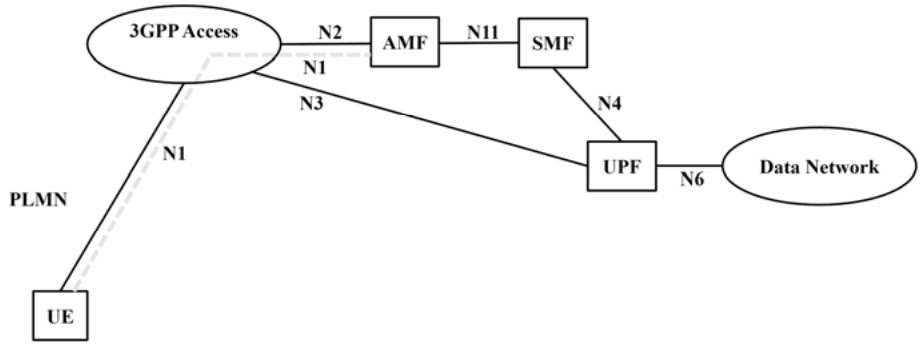

Fig. 9.7: 3GPP access with NR technology [38]

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].

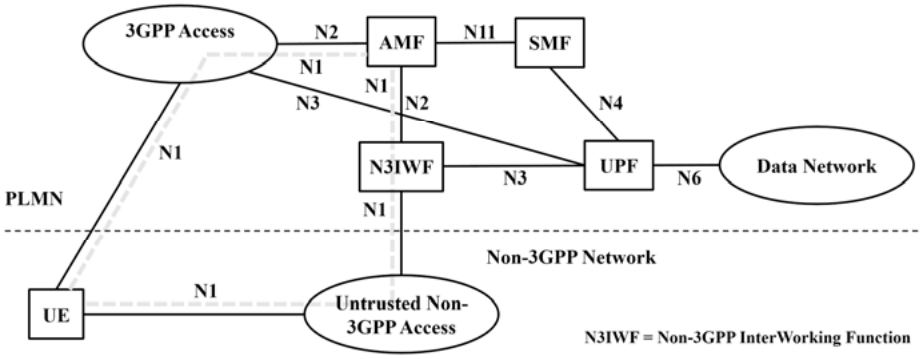

Fig. 9.8: Untrusted Non-3GPP access, e.g., with WLAN [38]

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].

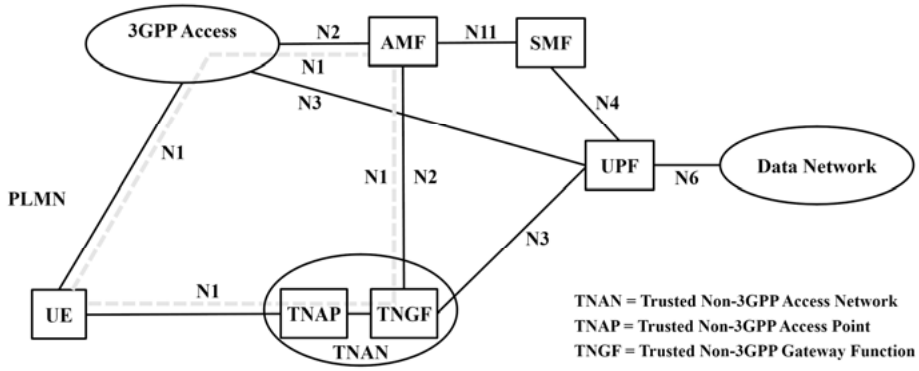

Fig. 9.9: Trusted Non-3GPP access, e.g., with WLAN [38]

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.

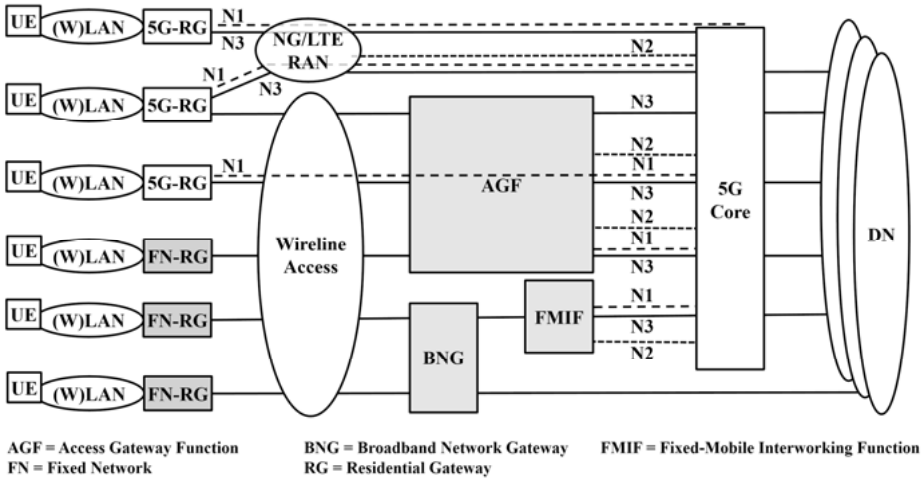

Fig. 9.10: Access network for Fixed Mobile Convergence (FMC) with 5G core [86]

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 are 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].

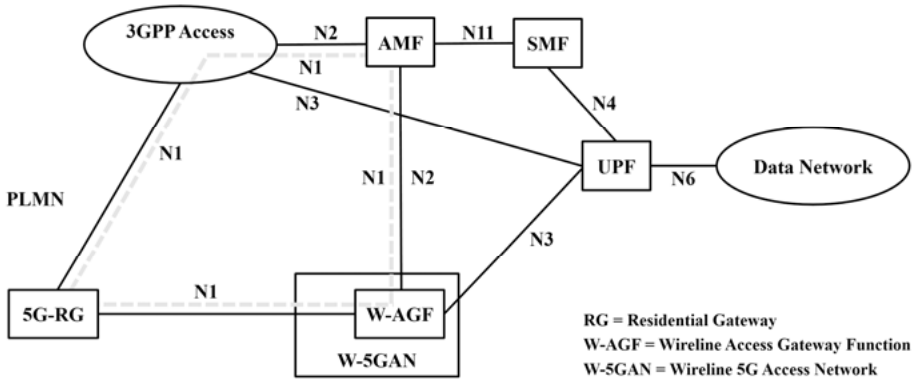

Fig. 9.11: Wireline 5G access network for 5G-RG [38]

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].

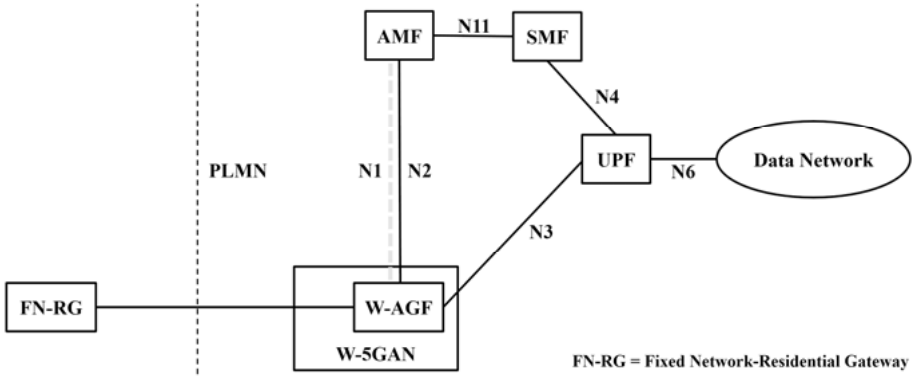

Fig. 9.12: Wireline 5G access network for FN-RG [38]

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].

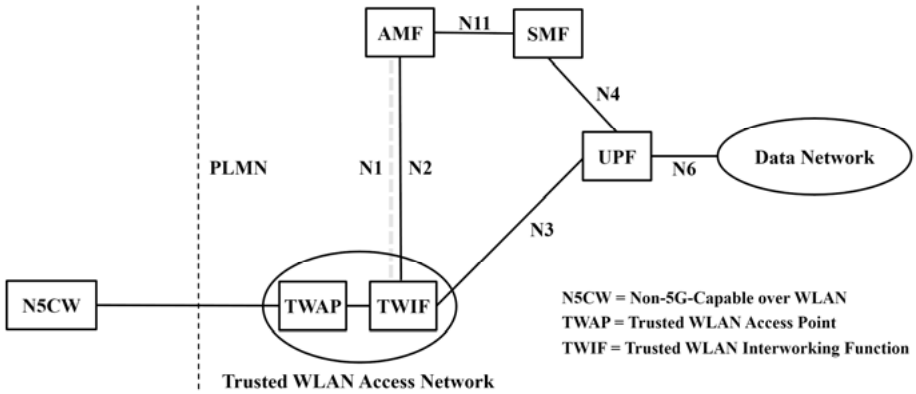

Fig. 9.13: 5G access network for N5CW-Geräte [38]

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 but in Release 16 only optional connection variant with 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].

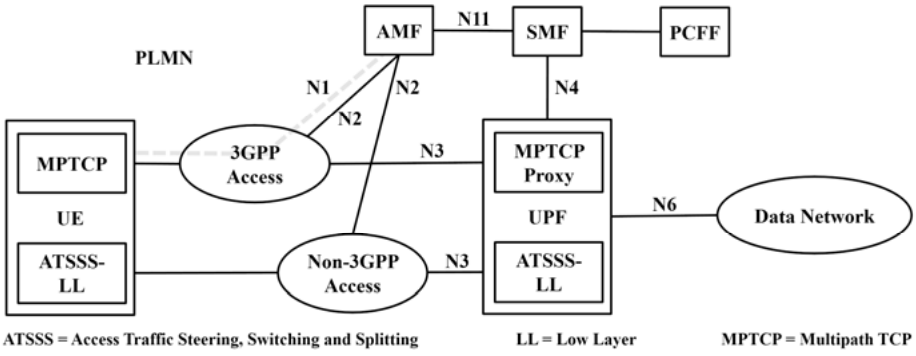

Fig. 9.14: 5G system with ATSSS support [38]

As mentioned above, 5G will include not only terrestrial radio and wireline access interfaces but also satellite-based access to achieve full convergence. The NG-RAN can also be hosted by an Unmanned Aerial System (UAS) with an integrated High Altitude Platform Station (HAPS) instead of a satellite. Carrier units for the latter systems can be uncrewed airships or 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.

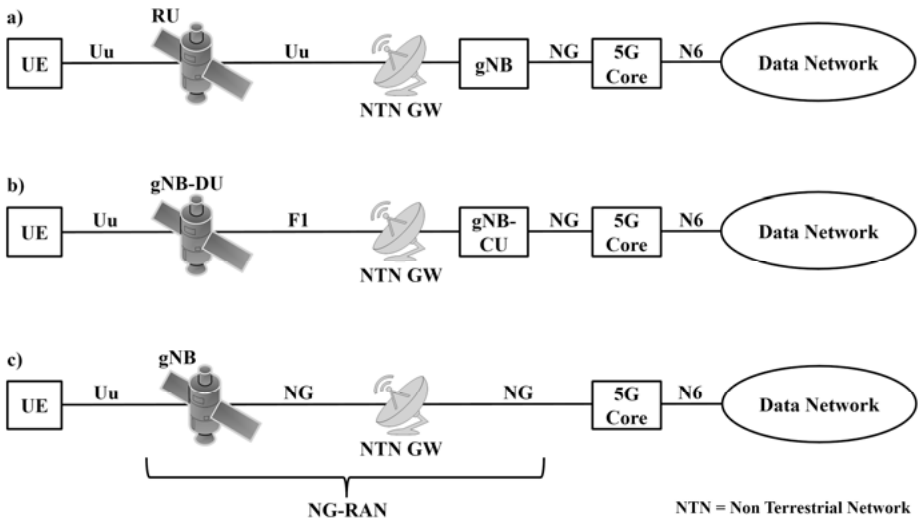

Fig. 9.15: Possible variants for 5G RAN satellite access [52]

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 pro-

vides 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 System in an Overall View

In summary, for this chapter, the following is an overall view of a 5G system. Based on the explanations on 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.

Figure 9.16 first structures the 5G network 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 Section 4.2) with their own frequencies, e.g., in Germany in the 3.7 to 3.8 GHz range (see Section 5.3).

The hardware in the core network consists of SDN switches and/or high-performance routers with dedicated HW, and 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.

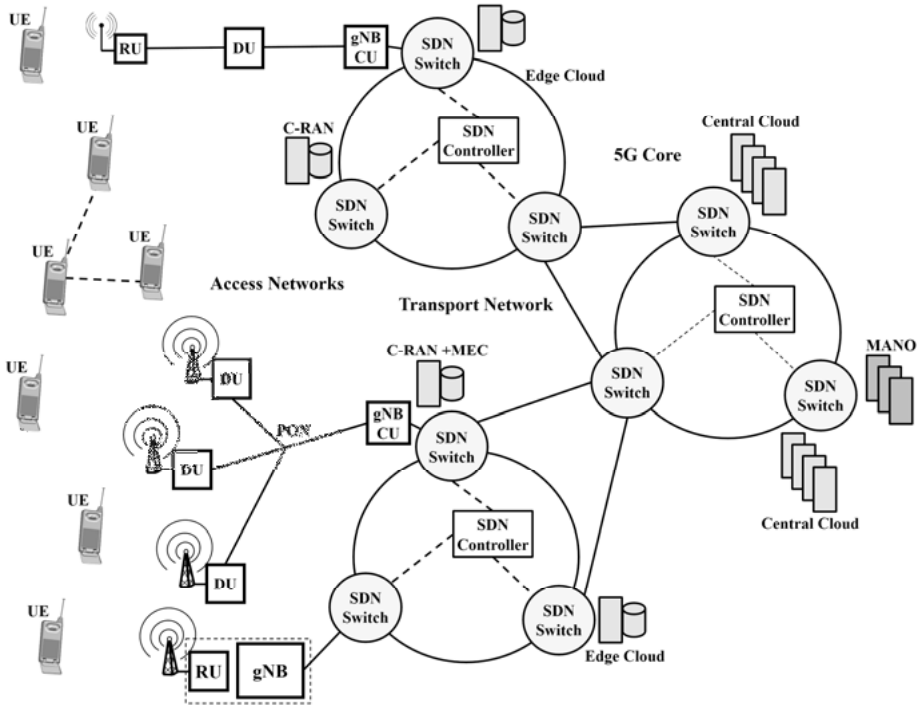

Fig. 9.16: 5G overall network view from the perspective of the physical layer [126]

This L1 architecture, as described above, is the basis for the functions of all higher layers in a 5G system, as shown in Figure 9.17. 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 8.2) and supports modularity, flexibility, and the building of network slices (see Section 8.3). 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.

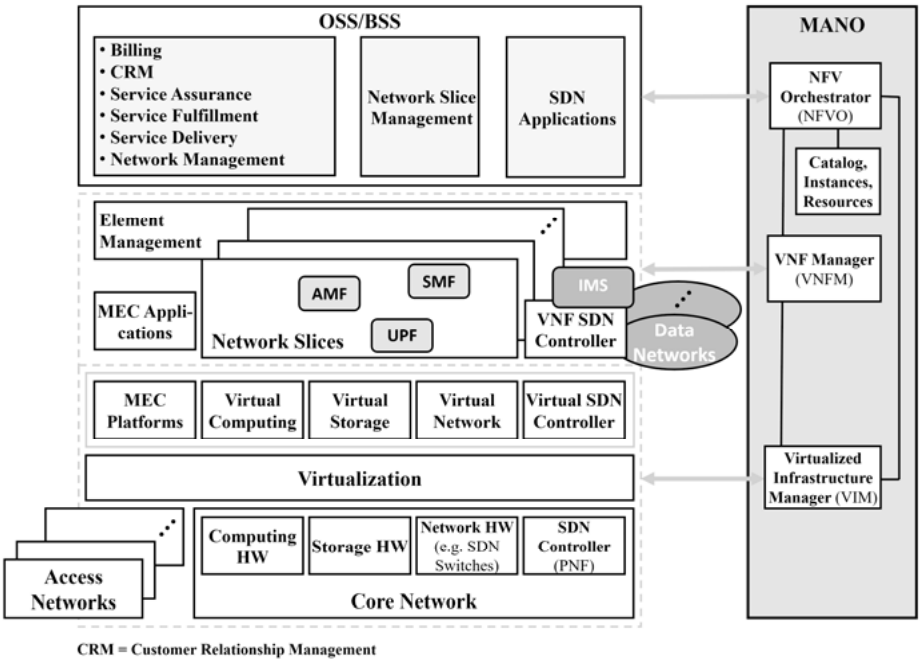

Fig. 9.17: 5G system in an overall view

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).

In conclusion to 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 9.16 and 9.17 (see Chapter 9, Section 2.5 and Figure 2.29).

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, which have 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 9.17. 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.

https://doi.org/10.1515/9783110724509-010

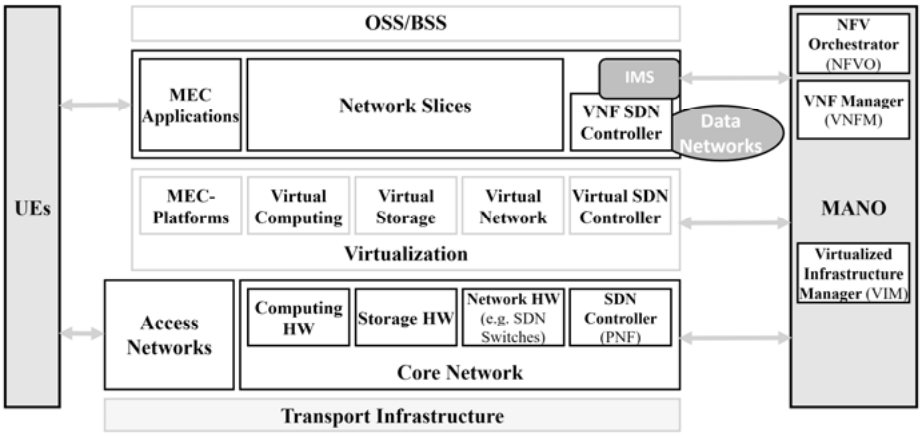

Fig. 10.1: Security-relevant components, subsystems, infrastructures, platforms, and functions of a 5G system

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].

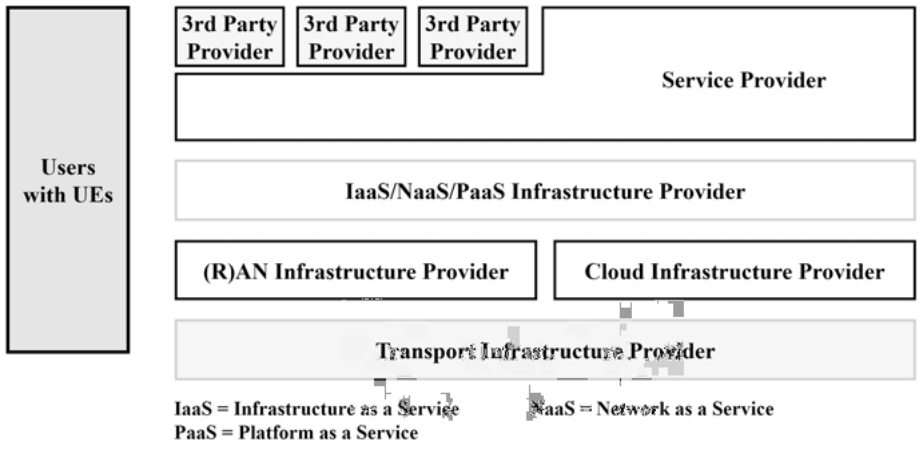

Fig. 10.2: 5G system with multiple actors [88]

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.

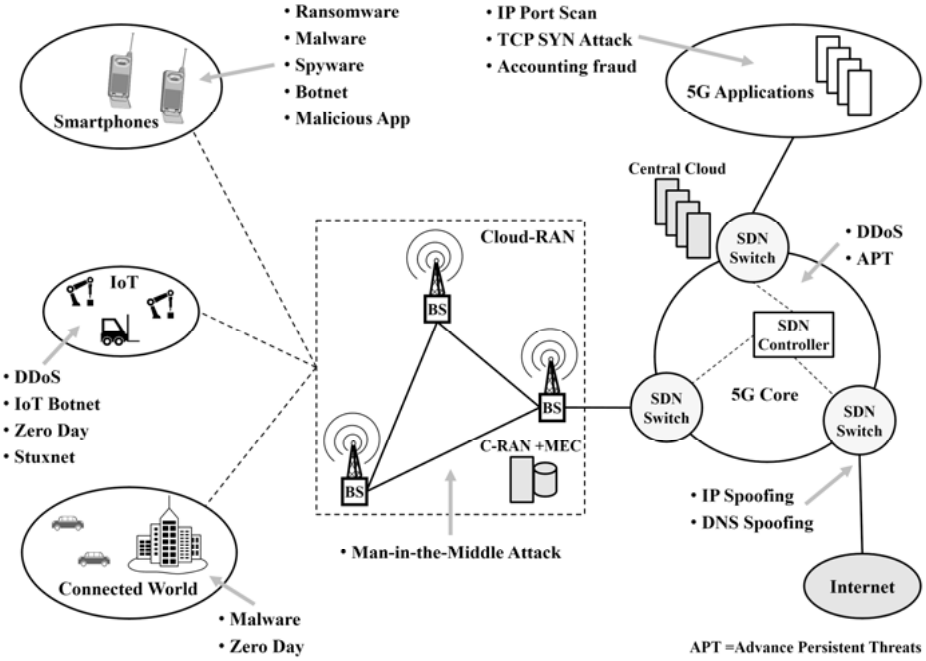

Fig. 10.3: Examples of threats on a 5G network [125]

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 will discuss each of these three security areas in the following three sections.

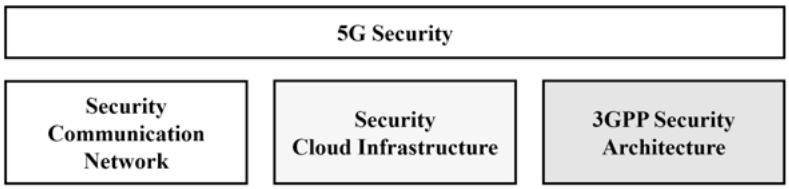

Fig. 10.4: Three pillar model of the 5G security framework [159]

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) [84; 205] 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 multi-stage filters and adaptive controls (mitigation devices) to limit damage
- Detection of botnets
- Prevention of manipulation of inter-domain routes
- Analysis of traffic to detect attacks or errors and to take appropriate protective measures
- Monitoring infrastructure to continuously identify and prevent threats
- Also, monitor the network concerning infected systems of customers
- Cooperation with anti-malware vendors
- Authentication
- Ensuring the availability of emergency numbers
- Encrypted transmission of VoIP data
- 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; 205]:

- 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)
- The trustworthiness of manufacturers or suppliers of critical components must be investigated and proven by an appropriate declaration of the source of supply.
- 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
- 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. In particular, critical network functions and network elements should do not depend on a single supplier. This could also be achieved by using open standards like Open RAN.

To ensure secure operation according to these 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].

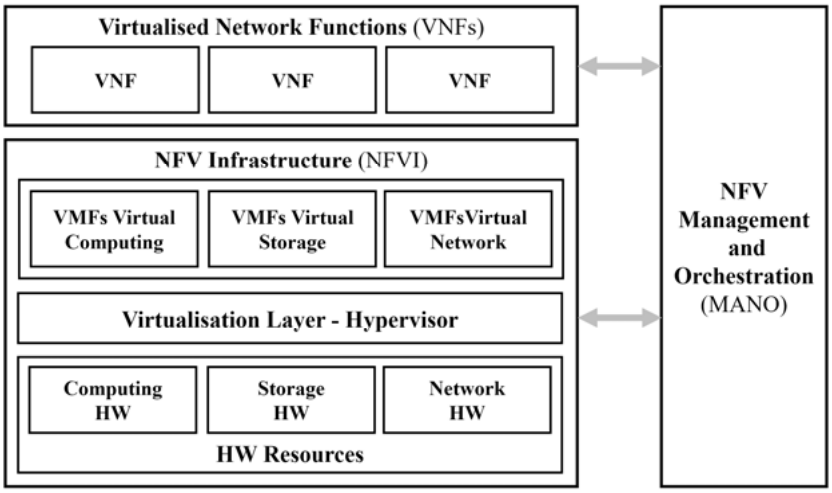

Fig. 10.5: NFV framework

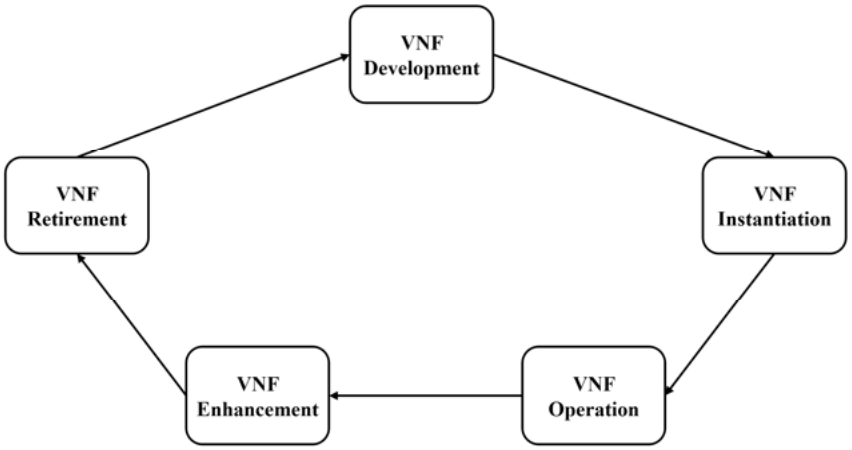

Fig. 10.6: VNF lifecycle [160]

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 (Northbound 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 then 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 operation 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 gateways 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].

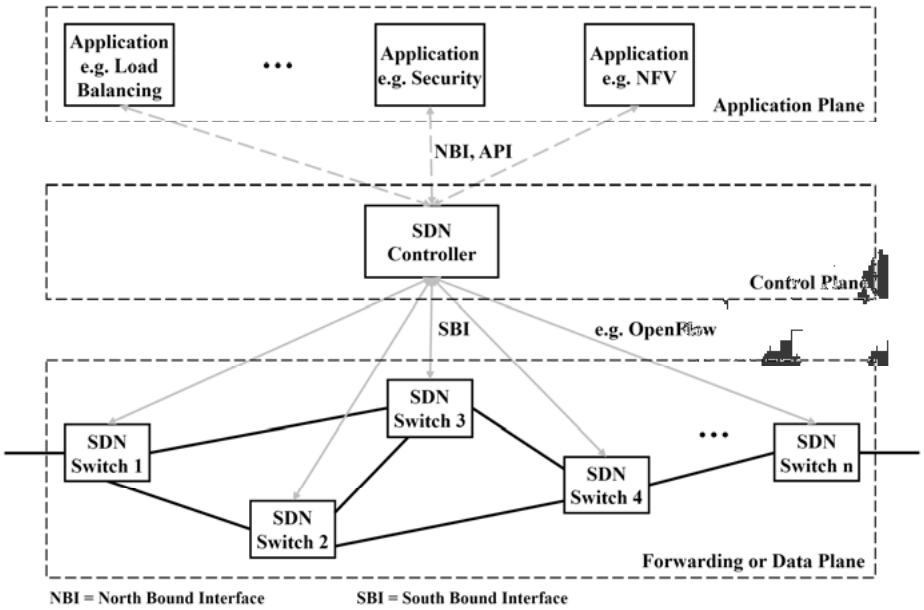

Fig. 10.7: SDN architecture

Network slices (see Section 8.3) based on the Service Based Architecture (see Section 8.2), 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.

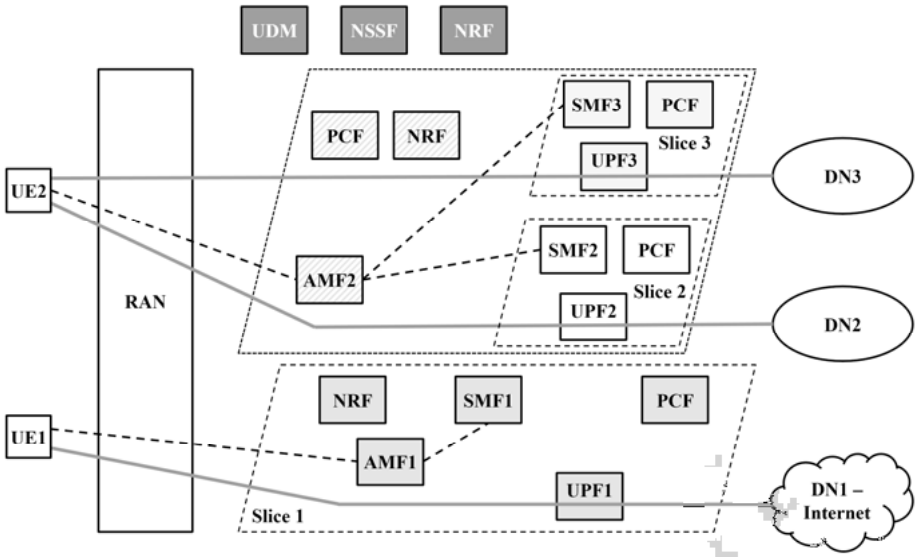

Fig. 10.8: Network slices

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 pishing, fraud, or exploitation of bugs
- Malicious insiders, i.e., current or former employees of the cloud infrastructure operator
- Advanced Persistent Threats (APTs)
- 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.

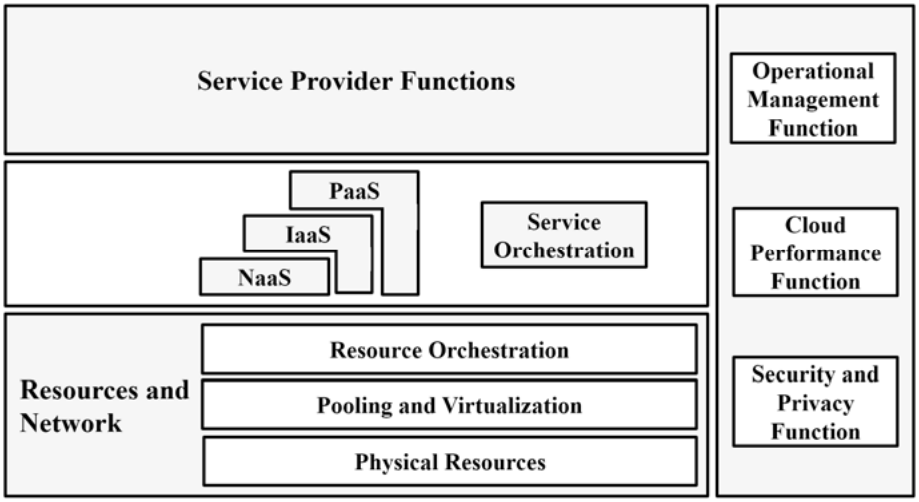

Fig. 10.9: Cloud environment

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 uses 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.

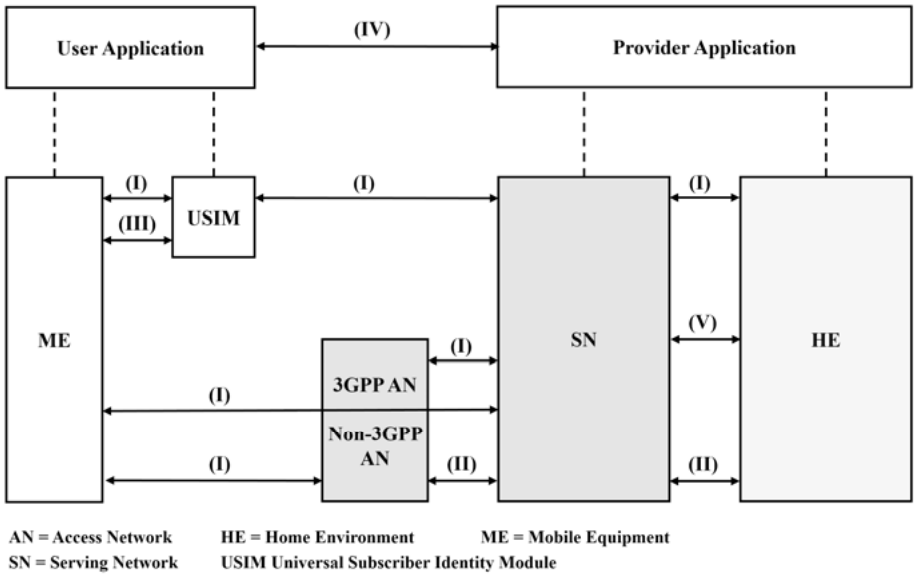

Fig. 10.10: 3GPP security architecture [45]

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].

Security requirements for 5G network elements and 5G network functions (see Chapter 8) are summarized in Table 10.2 [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	– 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 – 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	– 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 F1 interface when splitting a gNB into CU and DU – Secure transmission on E1 interface when dividing a CU into CU-CP and CU-UP
AMF (Access and Mobility Management Function)	– 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)	– Provides the authentication function via the AMF in the Serving Network – Supports the primary authentication of the UE
UDM (Unified Data Management)	– 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

5G network elements or 5G network functions	Security requirements
SIDF (Subscription Identifier Deconcealing Function)	– Responsible for resolving the SUPI (Subscription Permanent Identifier, the unique ID on SIM card) from the SUCI
AUSF (Authentication Server Function)	– 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	– Creation of trust zones, in any case between different 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)	– 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)	– 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
SEPP (Security Edge Protection Proxy)	– 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

Based on the requirements mentioned above, we can now provide an overview of the 3GPP security architecture implementation. Figure 8.6, from the description of the SBA approach to roaming in Section 8.2, is used here. If we take all network functions with security tasks from Table 10.2 and show them in grey in Figure 8.6, we will get Figure 10.11, which we will use for an introduction to the security procedures in a 5G system [159].

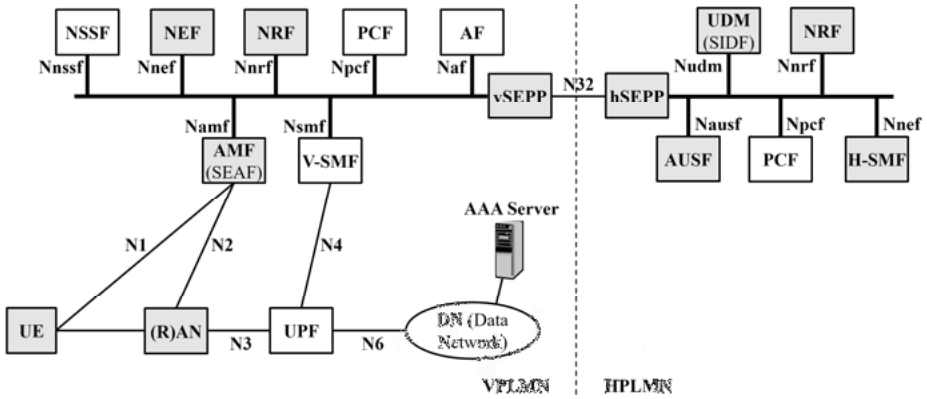

Fig. 10.11: 5G systems with SBA and security features for the case of roaming [37]

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, 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].

11 5G and Environment

In this chapter, we will take a look at the environmental impact of the 5G rollout. This includes the question of the influence of non-ionizing radiation due to radio transmission, energy consumption, and the need for raw materials.

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 already 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, free or released low-frequency spectral ranges are also used for 5G. 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 beamforming, the higher attenuation is compensated, and interference in the radio cell and from neighboring cells is reduced. 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 used (see Section 5.1). This will be 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.

https://doi.org/10.1515/9783110724509-011

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 that there will be a clear trend towards micro and picocells for 5G, especially in cities, to be able 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 [201].

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 very current and comprehensive summary of this topic can be found in [92; 185]. 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 stu-

dy 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; 185]

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 respect, the results of a metastudy [164] based on 94 publications for the frequency range 6 to 100 GHz and possible health effects are not very meaningful, and further research is needed.

"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." [185] According to [92], these are, among others:

- There are already many studies on the biological effects of HF-NIR 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.

The Federal Office for Radiation Protection (Bundesamt für Strahlenschutz, BfS) in Germany [78] does not assume adverse health effects of 5G after present scientific knowledge conditions but also sees still open questions. “In a further expansion step, higher frequency bands in the milli- or centimetre-wave range are also planned for 5G (e.g. in the 26 GHz, 40 GHz band or at up to 86 GHz). It can be assumed that no health effects are to be expected in these areas below the existing limit values. However, because only a few results are available for this area, the Federal Office for Radiation Protection still sees a need for research in this area.” “Open questions also arise from the fact that more transmitters are needed as data transmission volumes increase. This is not a 5G-specific problem – even today, ‘small cells’ are used in places with high user density. However, with the introduction of 5G, this will continue to increase. Although these ‘small cells’ will have a lower transmitting power, they will also be operated closer to places where people spend a considerable amount of time. It is not yet possible to estimate exactly how this will affect the extent to which the population is exposed to radiation. However, it can be assumed that the range of possible exposures will increase.” “Regardless of 5G, there are still scientific uncertainties regarding possible long-term effects of intensive mobile phone use.” [79] Besides, reference is made to a compilation of the Scientific Service of the European Parliament [119], which is interesting in this context and highlights various scientific views and results.

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 ^{0}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 mac-

ro 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].

Figure 11.1 shows network architectures with different cell types. 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. In these situations, both the base station and a mobile terminal have to 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 could also be offered 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.

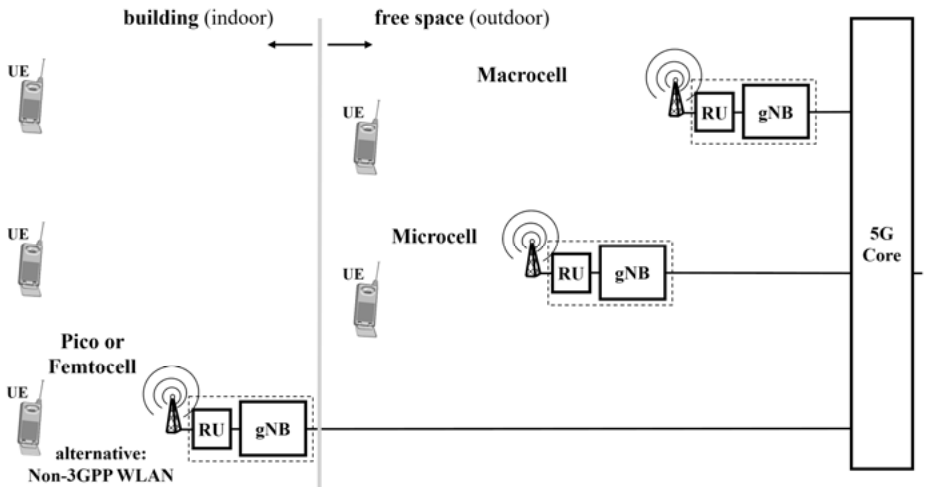

Fig. 11.1: Influence of the network architecture with different types of radio cells on NIR immissions

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, corresponding 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 and Raw Materials

In the past, the consumption of electrical energy by a mobile network accounted for 15% or more of operating costs. Of these, the RAN, i.e., the base stations, accounted for around 80%. Although 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 currently still have a significantly higher energy consumption than normal systems with passive antennas. However, the vendors all state that this can be lowered with the next generations. 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. The result is, according to [108], 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.

It is therefore 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, can be implemented.

As explained in Section 9.4, 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 the introduction of 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 CO^2 emissions, we have to consider counteractive measures. This can be, on the one hand, the increase of energy efficiency, e.g., by improvements in hardware, on the other hand, the use of renewable energy for the power supply of the 5G systems, not only in operation but also in production. Besides, it would also help to switch off old, energy-inefficient 2G and 3G technology [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].

12 Future Developments

As stated in Section 5.2 on standardization, 3GPP Release 16 for 5G will be fully standardized by the end of 2020. A 5G system compliant with Release 16 will then correspond to the IMT-2020 target system defined by the ITU (see Sections 4.3 and 5.2). The next step in the further development of 5G will be – as already mentioned in Section 4.3 and Section 5.2 on standardization Release 17. But this is certainly not the end of the 5G evolution; Release 18 has previously been announced. 3GPP Releases 17 and 18 are, therefore, discussed in more detail in Section 12.1. 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 not yet an issue in standardization, but of course, in research. Section 12.3 summarizes the first results.

12.1 Further Development of 5G

As already mentioned, the next steps in the further development of 5G will be the 3GPP Release 17 and implementations of 5G systems based on it. The most crucial service requirements, which complement Releases 15 and 16, are summarized in Table 12.1 [23]. Particular mention should be made here of multiple access technologies, IoT mass operations, cross-network mechanisms, e.g., for disasters, services for UAVs, audiovisual productions, and medicine.

Tab. 12.1: Additional service requirements for 5G system in 3GPP Release 17 [23]

Requirements "Basic capabilities"
Multiple access technologies, e.g., NG-RAN, WLAN, fixed broadband access network, 5G satellite access. Simultaneously use of two or more access technologies by a UE
Efficient IoT mass operation with up to 1000000 connections/km^2
Priority services,e.g., for multimedia priority service, emergency, medical, public safety
Indirect UE-5G network connection via several serially linked relay UEs (several hops) or parallel indirect connections, using various RAN techniques. E.g., for smart home, farming, factory, public safety
APIs to allow a third-party to customize a dedicated physical or virtual network or a network slice
Efficient network resource utilization by using information gathered by sensors, the utilized access technologies, the application context, and the application traffic characteristics
Wireless self-backhaul using NR and E-UTRA
Extensions for the provision of flexible broadcast/multicast services
Dynamic subscription generation and management for IoT

https://doi.org/10.1515/9783110724509-012

Requirements “Basic capabilities”
Support of an energy-saving mode by the 5G access network
Extreme long-range coverage (up to 100 km) in low-density areas (up to 2 user/km^2)
Services from more than one network simultaneously on an on-demand basis
Selection among any available PLMN/RAT combinations
Enhancements for eV2X
NG-RAN sharing
Unified access control
Real-time E2E QoS monitoring
5G LAN-type service
Non-public 5G networks
Positioning services
Support of cyber-physical control applications in the vertical domains of factories of the future, electric power distribution, central power generation, and rail-bound mass transit
Steering of roaming from the home network to register UE in another network, e.g., due to higher priority of the then visited network due to business arrangements
Minimization of service interruption in the event of a disaster (e.g., fire) by using the services (e.g., telephony) of another network in the affected area, roaming of the UE
Control of and provision of services (e.g., video) for UAVs (e.g., drone), including operation of the onboard radio interfaces
Video, images, and audio for professional applications such as audiovisual productions (e.g., in radio and television studios, at sporting 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
Very high availability of > 99,9999% for IoT traffic at up to 1000 UEs/km^2, etc.
High data rates and low end-to-end latency: ≤ 1 Gbit/s, 5 ms for VR from cloud/edge; < 1 Gbit/s, 10 ms for Gaming, etc.
For indirect UE-5G network connection via relay UE: ≤ 1 Gbit/s, 10 ms for 50 UEs/house; ≤ 5 Mbit/s, 50 ms for 10000 UEs/factory, etc.
High accuracy positioning: up to 0,2 m horizontal, 0,2 m vertical
Security requirements
Charging requirements

In 2020, the performance features and functionalities will be defined beyond the scope of Release 16. Table 12.2 provides an overview that has not yet been finalized. Particularly noticeable here are the new topics:

– Frequency range extension in RAN to 52,6 – 71 GHz
– Consideration of Non Terrestrial Networks (NTN), including satellites and UASs (Unmanned Aerial Systems) (see Section 9.3)

– Multicast/Broadcast via NR RAN
– Direct UE-UE communication (NR side link) including forwarding (relay)
– ATSSS (Access Traffic Steering, Switching and Splitting) (see Section 9.3), etc.

Tab. 12.2: Features and functionalities of 3GPP Release 17 [62; 61; 152]

Features
New
Frequency range extension to 52,6 – 71 GHz
NR via NTN with satellites and UAS (Unmanned Aerial System)
Multicast at NR
ATSSS (Access Traffic Steering, Switching and Splitting)
Network Slicing phase 2
Advanced Interactive Services with specific requirements for latency, availability, and bit rate
Multimedia Priority Service (MPS) for prioritized communication of emergency organizations and security authorities
Study on eXtended Reality (XR) via NR (combines virtual and real-world, e.g., for human-machine interaction)
Support of UEs with Multi SIM cards
Support of UEs with low performance or low complexity
Network automation phase 2
Improvements and enhancements
Forwarding direct UE-UE communication at NR (NR side link relay)
NR MIMO
Dynamic Spectrum Sharing (DSS) of a frequency range for 5G or LTE according to user needs
Multi-Radio DC (Dual Connectivity)
Network coverage with NR for FR1 and FR2
Position accuracy with NR
Integrated Access and Backhaul (IAB) for topology changes, simultaneous operation of access and backhaul links, and routing
RAN slicing
Fixed Mobile Convergence
LAN services via 5G
V2X services
URLLC for IoT in industrial environments
Handling of short data packets at mMTC (small data), e.g., from sensors
Saving electrical power for UEs
Non-public networks
SON/MDT (Self-Organizing Networks/Minimization of Drive Tests)

Figure 12.1 shows the timetable for the standardization of Release 17 for 3GPP, with the standards being finalized by spring 2022.

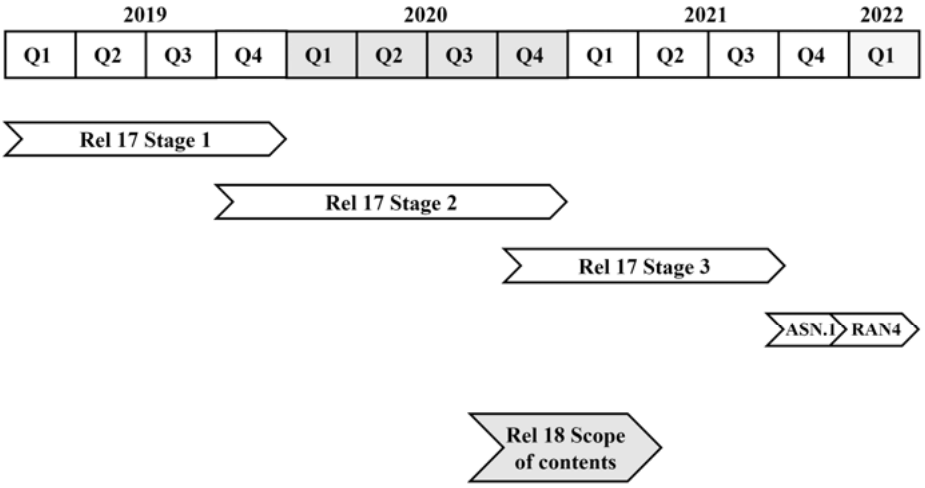

Fig. 12.1: Time schedule for the standardization of 3GPP Release 17 [58; 62]

Figure 12.1 also indicates the subsequent Release 18. The desired features and functionalities as the basis for standardization will be defined in 2021. Further details are not yet known.

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 is organized in three subgroups for:

- 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] [99].

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].

[202; 203] 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

From the mentioned more detailed requirements, the requirement categories are evaluated relatively with scores between 1 (not important) and 10 (extremely important). Table 12.3 shows the results for the twelve Network 2030 use cases [202; 203].

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.

Tab. 12.3: Network 2030 use cases and evaluated requirements [202; 203]

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 (SIoT): 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

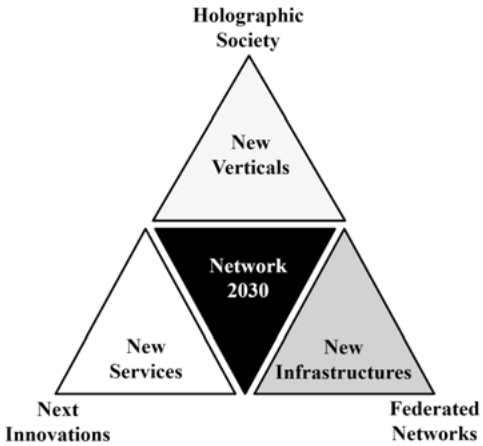

Fig. 12.2: Network 2030 vision [123]

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**	**Holografic 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 operational 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 [204]:

- 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).

Figure 12.4 shows the application scenarios considered for a Network 2030:
– 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 application scenarios eMBB, URLLC, and mMTC in Section 4.1 and especially in Figure 4.1.

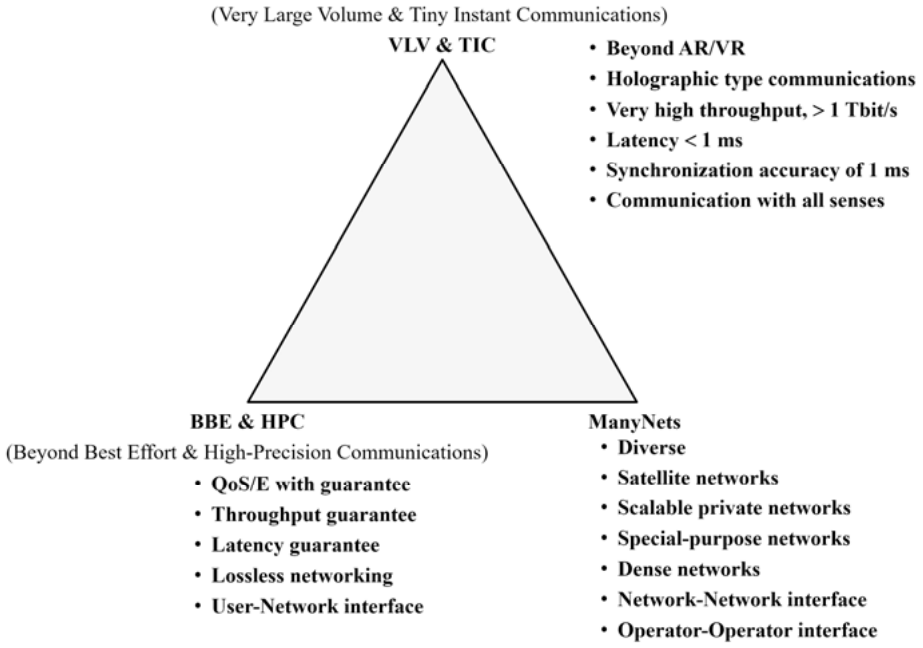

Fig. 12.4: Application scenarios for a Network 2030 according to ITU-T FG NET-2030 [124]

In a first approach [204] 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 programmatically 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 micro (i.e., IoT devices) and new advanced together with existing devices.
- Heterogeneity in communication, computing, storage, service, and their integration
- Native slicing
- Unambiguous naming 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 meeting 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.

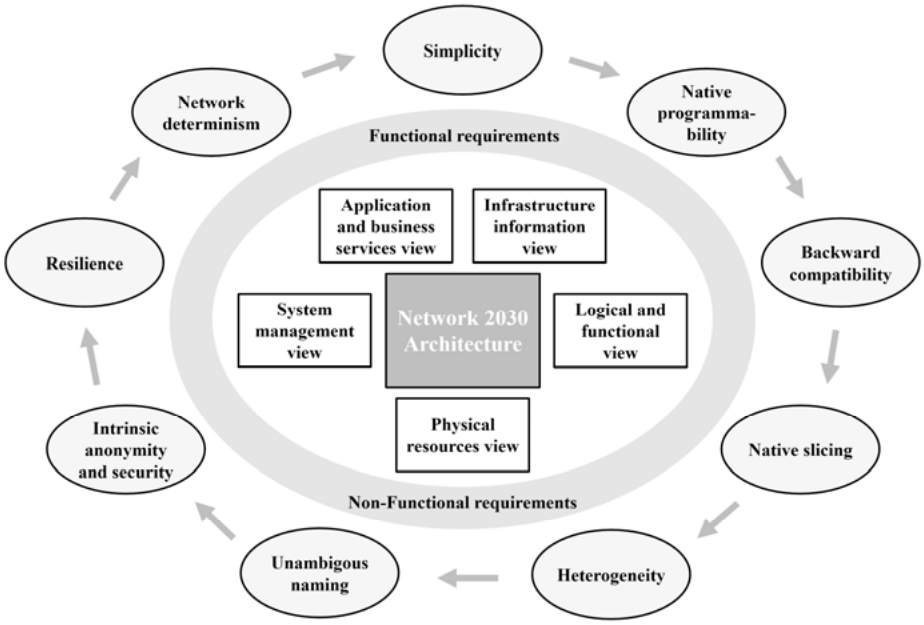

Fig. 12.5: Network 2030 architecture principles [204]

[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, [204] 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. However, the corresponding code rate corresponds to the access bandwidth to 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 [204].

12.3 6G Considerations

Apart from the activities of FG NET-2030 (see Section 12.2), which refer to a general network or Internet for the period after 2030, the considerations on 6G are all still in the research stage. An example of bundled research on 6G is the 6G Flagship Project [67] at the University of Oulo in Finland. Initiated by this project, the first 6G Wireless Summit conference took place in 2019 [68]. Their results were summarized in a document with 6G research visions [122], which will be updated in the future. In addition, there are already numerous scientific publications, but not all of them are based on the same assumptions, and they come to slightly different results and visions for 6G. However, a red line does appear to be visible, although with relatively large uncertainties. In the following, we will follow this line to give an idea of the applications that make a 6G system necessary, the requirements for 6G, a possible network architecture, and further technologies.

Based on [183], Figure 12.6 shows application scenarios and requirements that significantly exceed 5G:

- Holography applications in communication
- Communication with all senses - hearing, seeing, feeling/touching, tasting, and smelling
- Super HD (SHD) and Extreme HD (EHD) video,

all three summarized under FeMBB (Further-enhanced Mobile Broadband).

- Industrial Internet with delay times in the range of 10 μs
- Enhancements or improvements in support of IoT applications and autonomous vehicles,

subsumed under ERLLC (Extremely Reliable and Low-Latency Communications).

- Internet of Nano-Things and Internet of Bodies with wearables and implants at extremely low power consumption,

with the heading ELPC (Extremely Low-Power Communications).

– Underwater and
– Space communication
– Service usage on hyper-high-speed trains,

characterized as LDHMC (Long-Distance and High-Mobility Communications).

– Significant enhancements to mMTC or the Internet of Everything (IoE)

under the keyword umMTC (Ultra-Massive Machine-Type Communications).

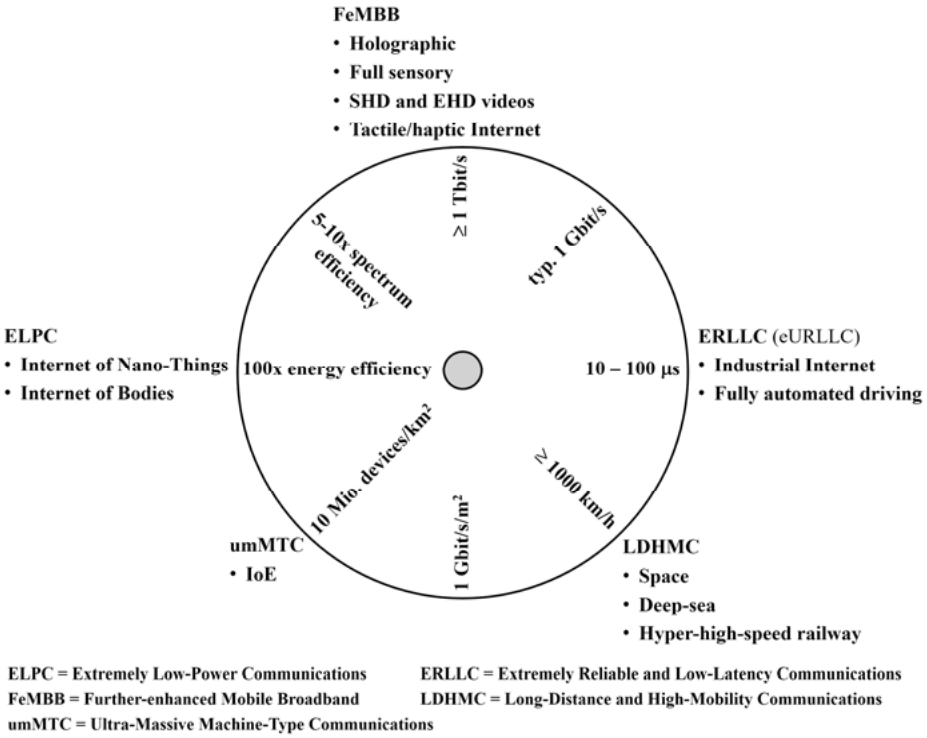

Fig. 12.6: Application scenarios and requirements for 6G according to [183]

The Japanese network operator NTT pursues a slightly different, more general approach to 6G. They assume four areas in which 6G communication is supposed to help considerably [146]:

– Solving social problems: They expect that telework supported by 5G, remote control of machines/processes, telemedicine, distance learning, and autonomous systems will already promote regional development and counteract low birth rates, aging societies, and labor shortages. 6G will provide access to people, information, and goods worldwide with a user experience that is so real that the geographical location becomes less important. Thus the social and cultural differences between urban and rural areas increasingly disappear, and the

problems that have been associated with them until now fade into the background.
- Communication between people and things: By supporting new functions such as VR, AR, extremely high-resolution images, holography, and communication with all senses, the interaction between people and things leads to realistic experiences. Time and, above all, space restrictions for the use of business services, participation in games, watching sports events are becoming less relevant. The popularity of IoT services will increase significantly. The requirements for bit rates, latency, and processing power will rise accordingly.
- Expansion of the communication environment: Communication is considered to be as important as water and power supply. Therefore, no environment-specific settings or application restrictions should be required for ease of use. Skyscrapers, drones, flying cars, airplanes, and even outer space become regular activities and communication environments. Also, the need for communication at sea and underwater will increase.
- Extended fusion of cyberspace and physical space: This already begins with 5G, but it is only with 6G that physical and cyber worlds can be merged or mapped to each other in real-time via sensors and actuators and the exchange and processing of huge amounts of data due to the extremely short delay times and the extremely high bit rates.

These considerations lead in [146] to the requirements summarized in Figure 12.7 with the six areas:
- Extremely high data rate and traffic capacity
- Extremely low latency
- Extremely high availability
- Extremely high connection density
- Extremely low energy consumption (also with power supply via radio) and very low costs
- Extreme network coverage, e.g., even in space.

The summary of the results of the first 6G Summit 2019 in [122] leads to very similar results as in [183] and [146]: Peak data rates up to 1 Tbit/s, delays down to 0.1 ms, up to 20 years battery life, 100 devices per m^3, 10000 times more traffic compared to 5G, 99.9999% availability, 10 times more energy efficiency compared to 5G, and positioning accuracy of 10 cm indoors and 1 m outdoors [122]. The above-mentioned source [183] confirms or even tightens the requirements for 6G and compares them with the requirements for an IMT-2020 or 5G system shown in Figure 4.6. Figure 12.8 shows this comparison.

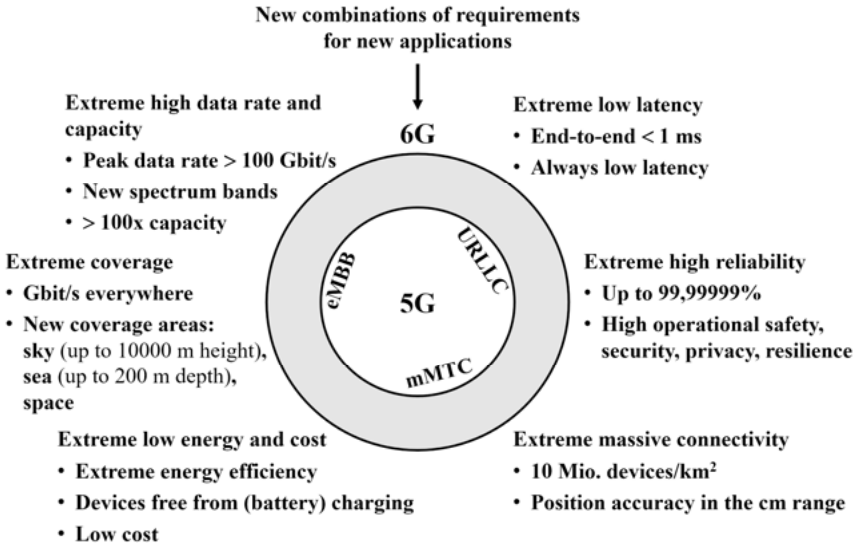

Fig. 12.7: Requirements for 6G according to [146]

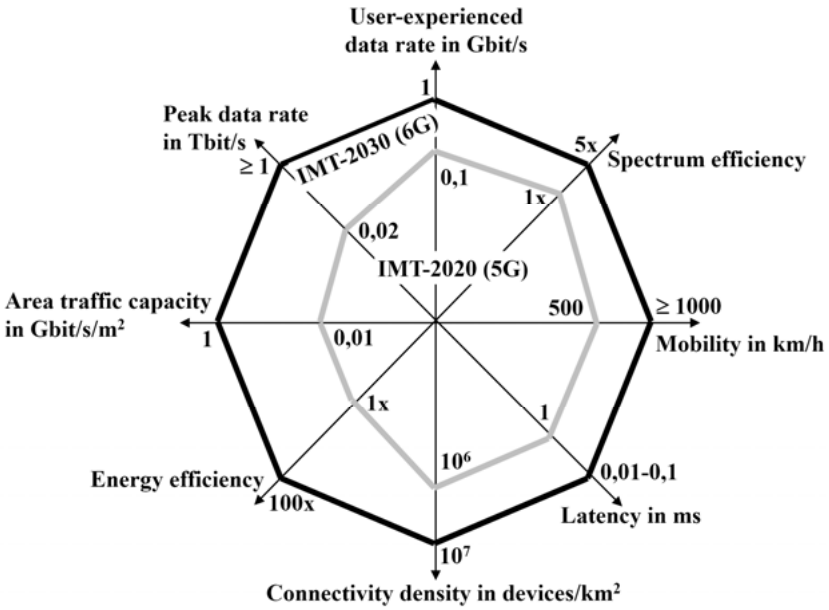

Fig. 12.8: Most essential requirements for a 6G compared to an IMT-2020 system [183]

In the various sources on possible future 6G solutions, there is a broad agreement not only on the requirements but also on the potential technologies. However, some

see the inclusion of underwater communication; others do not even mention it. On the other hand, there is agreement on the necessity of space communication.

It is, of course, too early to make any concrete statements about a 6G network architecture and the techniques and technologies that will be used. Nevertheless, the following is an attempt to paint a picture of the future of 6G to give the first impression. [183] and [107] assume a network that supports communication with fixed lines, terrestrial radio, radio in space, and underwater. This would be a multidimensional network that works largely autonomously due to the extensive use of artificial intelligence. The UAS or UAV (Unmanned Aerial Vehicle) and HAP (High Altitude Platform) in near-earth space already mentioned in Section 9.3, as well as satellites such as LEO and GEO types in space, are used for this purpose. Figure 12.9 shows an example of such a multidimensional 6G network [107].

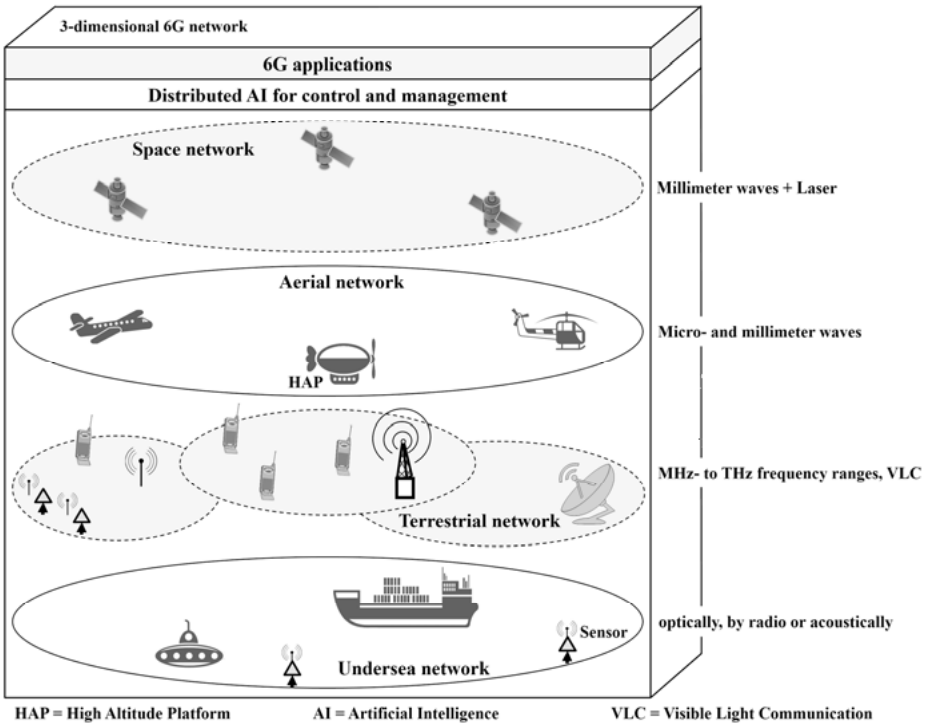

Fig. 12.9: Illustrative example of a multidimensional 6G network [107]

In space, millimeter waves and laser transmission are used. UAVs and HAPS can use low-frequency microwaves and also millimeter waves. Terrestrial radio networks will use the existing frequency ranges, such as 5G, plus higher-frequency millimeter waves. Because of the extremely high data rates required in the Tbit/s range (for

local use only), THz spectra from 0.1 to 10 THz and more are also to be used. Alternatively, communication with light, known as Visible Light Communication (VLC), could be applied here, especially in buildings, using LEDs (Light-emitting Diode). A comparison of the two latter transmission technologies is provided by Table 12.4 [107].

Tab. 12.4: Comparison of transmission with radio in the THz range and with visible light [107]

	THz frequency range	**Visible light**
Available bandwidth	up to hundreds of GHz	hundreds of THz
Transmission	possible without line of sight (Non Line of Sight, NLOS)	Line of Sight (LOS) required
Electromagnetic radiation	yes	no
Achievable data rate	> 100 Gbit/s	10 Gbit/s
Regulation	licensed	unlicensed
Interference	yes	no
Costs	high	low
Transmission power	high	low

For the underwater communication mentioned above, there are also different transmission techniques, with low-frequency radio, acoustically or optically with lasers. Table 12.5 shows a comparison of these [107].

Tab. 12.5: Comparison of transmission techniques, with radio, acoustic or optical, for underwater communication [107]

	Radio transmission	**Acoustic transmission**	**Optical transmission**
Attenuation	high	relatively low	depends on the turbidity of the water
Data rates	In the Mbit/s range	In the kbit/s range	In the Gbit/s range
Latency	medium	high	low
Transmission range	< 10 m	< 100 km	< 100 m
Power consumption	medium	high	low

In summary, Table 12.6 provides an overview of a conceivable 6G system and compares it with 5G and 4G. On the one hand, it becomes evident that there are extended application scenarios that have already been briefly explained above in the requirements: FeMBB, ERLLC, umMTC, LDHMC, and ELPC. These also cover new applications such as holographic communication. As far as the essential network

technologies are concerned, the main difference to 5G is the pervasive and distributed use of AI (Artificial Intelligence). The goal is comprehensive networking of people and things and this multidimensional. The KPIs will be a power and more above those of 5G. Due to the enormous demands on radio technology, e.g., to achieve significant coverage in the THz range, the following technologies might be necessary: SM-MIMO (Spatial Modulation-MIMO) with more than 10,000 antenna elements, LIS (Large Intelligent Surfaces) for phase change of electromagnetic waves, or HBF (Holographic Beamforming) for redirection of electromagnetic waves around objects, as well as OAM Multiplexing (Orbital Angular Momentum) for multiplexing electromagnetic waves using angular momentum. Table 12.6 also mentions the use of Blockchain technology for the efficient, decentralized, and secure management of the distribution or use of the same spectrum by several terminal devices (spectrum sharing). Furthermore, the use of quantum communication and quantum computing is also mentioned in the context of 6G [183].

Tab. 12.6: Network characteristics of 6G compared to 5G and 4G [183]

	4G	5G	6G
Usage scenarios	– MBB	– eMBB – URLLC – mMTC	– FeMBB – ERLLC – umMTC – LDHMC – ELPC
Applications	– HD videos – Voice – Mobile TV – Mobile Internet – Mobile Pay	– VR/AR/360^0 videos – UHD videos – V2X – IoT – Smart City/ Factory/Home – Telemedicine – Wearables	– Holographic – Tactile/haptic Internet – Full-sensory digital sensing and reality – Fully automated driving – Industrial Internet – Communication in space – Deep-sea communication – Internet of Bio-Nano-Things
Network characteristics	– All-IP	– Cloudification – Softwarization – Virtualization – Network slicing	– AI use (Intelligentization) – Cloudification – Softwarization – Virtualization – Network slicing
Service users/objects	People	People and things	People, things, and world
Peak data rate	100 Mbit/s	20 Gbit/s	≥ 1 Tbit/s
Experienced data rate	10 Mbit/s	100 Mbit/s	1 Gbit/s

	4G	5G	6G
Spectrum efficiency	1x	3x that of 4G	5-10x that 5G
Network energy efficiency	1x	10-100x that of 4G	10-100x that of 5G
Area traffic capacity	0,1 Mbit/s/m^2	10 Mbit/s/m^2	1 Gbit/s/m^2
Connection density	10^5 devices/km^2	10^6 devices/km^2	10^7 devices/km^2
Latency	10 ms	1 ms	10-100 µs
Mobility	250 km/h	500 km/h	≥ 1000 km/h
Technologies	– OFDM – MIMO – Turbo Code – Carrier Aggregation – Hetnet (Heterogeneous Network) – ICIC (Inter-cell Interference Coordination) – D2D – Unlicensed spectrum	– mm-wave communications – LDPC – Flexible frame structure – Ultradense Networks – NOMA (Non-orthogonal Multiple Access) – Cloud/edge computing – SDN/NFV/network slicing	– THz communications – SM-MIMO (Spatial Modulation-MIMO) – LIS (Large Intelligent Surfaces) und HBF (Holographic Beamforming) – OAM Multiplexing (Orbital Angular Momentum) – Laser and VLC – Blockchain-based spectrum sharing – Quantum communications and computing – AI/Machine Learning

Abbreviations

3GPP	3rd Generation Partnership Project
3GPP2	3rd Generation Partnership Project 2
5GAA	5G Automotive Association
5GC	5G Core
5GMF	Fifth Generation Mobile communications promotion Forum
5GPPP	5G Infrastructure Public Private Partnership

A

a	Attributes
AAA	Authentication, Authorization and Accounting
ABF	Application-aware data Burst Forwarding
ACIA	Alliance for Connected Industries and Automation
A-CPI	Applications-Controller Plane Inerface
AF	Application Function
AGF	Access Gateway Function
AGV	Automated Guided Vehicle
AI	Artificial Intelligence
AKA	Authentication and Key Agreement
ALG	Application Layer Gateway
AMF	Access and Mobility Management Function
AMR	Adaptive Multi-Rate
AN	Access Network
AN	Access Node
API	Application Programming Interface
APT	Advance Persistent Threat
AR	Augmented Reality
ARPF	Authentication credential Repository and Processing Function
ARQ	Automatic Repeat Request
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
AVP	Audio Video Profile

https://doi.org/10.1515/9783110724509-013

B

B2BUA	Back-to-Back User Agent
BAKOM	Bundesamt für Kommunikation
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
BICC	Bearer Independent Call Control
BNG	Broadband Network Gateway
BNetzA	Bundesnetzagentur
BRG	Broadband Residential Gateway
BS	Base Station
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
CaaS	Communication as a Service
CAP	CAMEL Application Part
CBRS	Citizens Broadband Radio Service
CCS	Common Channel Signaling
CDN	Content Delivery Network
Ce	Connection establishment
CER	Capabilities-Exchange-Request
CHF	CHarging Function
CN	Core Network
ComCom	Kommunikationskommission
COPS	Common Open Policy Service
CriC	Critical Communications
CP	Control Plane
CP	Cyclic Prefix
CPE	Customer Premises Equipment
CP-OFDM	Cyclic Prefix-OFDM
CPS	Cyber Physical System
C-RAN	Cloud-RAN

C-RAN	Centralized-RAN
CRC	Cyclic Redundancy Check
CRG	Cable Residential Gateway
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
Ct	Connection termination
CU	Control Unit
CW	Code Word

D

D2D	Device-to-Device
D/A	Digital/Analog
DC	Dual Connectivity
D-CPI	Data-Controller Plane Interface
DDoS	Distributed Denial of Service
DL	Down Link
DM-RS	Demodulation References Signal
DN	Data Network
DNN	Data Network Name
DRB	Data Radio Bearer
DSCP	Differentiated Services Code Point
DSL	Digital Subscriber Line
DSS	Dynamic Spectrum Sharing
DSS1	Digital Subscriber Signalling system no. 1
DTMF	Dual Tone Multi-Frequency
DT	Digital Twins
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 Identity Register
EHD	Extrem HD-Video
ELPC	Extremely Low-Power Communications
eMBB	Enhanced Mobile Broadband
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
E-UTRAN	Evolved-UTRAN
eV2X	Enhancement of Vehicle-to-Everything
EVS	Enhanced Voice Services
Ex	Exchange

F

FAST	Facilitate America´s Superiority in 5G Technology
FCC	Federal Communications Commission
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
GPRS	General Packet Radio Service
GR	GPRS Register
GRE	Generic Routing Encapsulation
GSM	Global System for Mobile Communications
GSMA	Groupe Speciale Mobile Association
GTP-U	GPRS Tunneling Protocol-User plane

H

HAP	High Altitude Platform
HAPS	High Altitude Platform Station
HARQ	Hybrid Automatic Repeat Request
HBF	Holographic Beamforming
HE	Home Environment
HEO	High Elliptical Orbit
HF	High Frequency
Hetnet	Heterogeneous Network
HLR	Home Location Register
HLS	High Layer Split
HPLMN	Home PLMN
HSD	Huge Scientific Data
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
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
IF	Intermediate Frequency
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
IoT	Internet of Things
IP	Internet Protocol
IPsec	IP Security
IS	Interim Standard
ISC	IMS Service Control
ISDN	Integrated Services Digital Network
ISG	Industry Specification Group
ISUP	ISDN User Part
ITS	Intelligentes Transportsystem
ITU	International Telecommunication Union
ITU-R	ITU-Radiocommunication Sector
ITU-T	ITU-Telecommunication Standardization Sector

J

J	Joule
JSON	JavaScript Object Notation

K

KPI	Key Performance 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
M3UA	MTP3 User Adaptation
MAA	Multimedia-Auth-Answer
MAC	Medium Access Control
MANO	Management and Orchestration
MAP	Mobile Application Part
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
Megaco	Media Gateway Control Protocol
MEO	Medium Earth Orbit
METIS	Mobile and wireless communications Enablers for the Twenty-twenty Information Society
MFV	Mehrfrequenzwahlverfahren
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
MME	Mobility Management Entity

mMTC	Massive Machine Type Communications
mmW	millimeter waves
MPEG	Moving Picture Experts Group
MPLS	Multiprotocol Label Switching
MPS	Multimedia Priority Service
MPTCP	Multipath TCP
MR	Multi-Radio
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
MSC	Message Sequence Chart
MSC	Mobile Switching Center
MTC	Machine-Type Communications
MTP	Message Transfer Part
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
NCC	Network and Computing Convergence
NE-DC	NR-E-UTRA Dual Connectivity
NEF	Network Exposure Function
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
NLOS	Non Line of Sight
NOMA	Non-orthogonal Multiple Access
NR	New Radio
NR-DC	NR-NR Dual Connectivity
NRF	Network Repository Function
NSA	Non-Standalone
NSH	Network Service Header
NSI	Network Slice Instance
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

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-Ionizing Radiation
OS	Operating System
OSS	Operations Support System
OTT	Over The Top
OVSDB	Open vSwitch Database Management Protocol

P

PaaS	Platform as a Service
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
PDN	Packet Data Network
PDP	Packet Data Protocol
PDSCH	Physical Downlink Shared Channel
PDU	Protocol Data Unit
PGW	Packet Data Network Gateway
PKI	Public Key Infrastructure
PLMN	Public Land Mobile Network
PNF	Physical Network Function
PON	Passive Optical Network
POP	Point of Presence
PRACH	Physical Random Access Channel
ProSe	Proximity-based Services
PS	Packet-Switched
PSTN	Public Switched Telephone Network
PT	Payload Type
PUCCH	Physical Uplink Control Channel
PUSCH	Physical Uplink Shared Channel

Q

QAM	Quadrature Amplitude Modulation
QoE	Quality of Experience
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying

R

R	Receive
RADIUS	Remote Authentication Dial In User Service
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Resource Block
RCF	Radio Control Function
reconly	receive only
Rel	Release
REST	Representational State Transfer
RESTful	Representational State Transfer
RF	Radio Frequency
RFC	Request for Comments
RG	Residential Gateway
RIC	RAN Intelligent Controller
RLC	Radio Link Control

RNC	Radio Network Controller
RRC	Radio Resource Control
RRH	Remote Radio Head
RTCP	RTP Control Protocol
RTP	Real-time Transport Protocol
RTR	Rundfunk und Telekom Regulierungs-GmbH
RU	Radio Unit

S

SA	Standalone
SAA	Server-Assignment-Answer
SaaS	Software as a Service
SAE	System Architecture Evolution
SAR	Server-Assignment-Request
SAR	Spezific Absorption Rate
SAS	Spectrum Access System
SBA	Service Based Architecture
SBC	Session Border Controller
SBI	South Bound Interface
SBI	Service-based Interface
SCC AS	Service Centralization and Continuity Application Server
SCCP	Signaling Connection Control Part
SCN	Switched Circuit Network
S-CSCF	Serving-CSCF
SCTP	Stream Control Transmission Protocol
SDAP	Service Data Adaptation Protocol
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-Gateway
SHD	Super HD Video
SI	Service Index

SIDF	Subscription Identifier De-concealing Function
SIGTRAN	SIGnalling TRANsport
SIoT	Socialized Internet of Things
SIP	Session Initiation Protocol
SL	Subscriber Locator
SLF	Subscription Locator Function
SM-MIMO	Spatial Modulation-MIMO
SMF	Session Management Function
SN	Serving Network
S-NSSAI	Single-NSSAI
SON	Self-Organizing Networks
SPI	Service Path Identifier
SRB	Signaling Radio Bearer
SRI	Satellite Radio Interface
SRVCC	Single Radio Voice Call Continuity
STIN	Space-Terrestrial Integarted Network
Sub	Subsciber
SUCI	Subscription Concealed Identifier
SUL	Supplementary Uplink
SUPI	Subscription Permanent Identifier
SW	Software

T

T	Transmit
TAS	Telephony Application Server
TCP	Transport Control Protocol
TDD	Time Division Duplexing
TDoS	Telephone Denial of Service
TIA	Telecommunications Industry Association
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

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
UAR	User-Authorization-Request
UAS	Unmanned Aerial 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 Telcommunications System
UP	User Plane
UPF	User Plane Function
URI	Uniform Resource Identifier
URLLC	Ultra-Reliable and Low Latency Communications
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
WiMAX	Worldwide Interoperability for Microwave Access
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, February 2008

[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

https://doi.org/10.1515/9783110724509-014

[20] TR 21.916 V0.5.0: Release 16 Description; Summary of Rel-16 Work Items (Release 16). 3GPP, July 2020
[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] TS 22.261 V17.3.0: Service requirements for the 5G system; Stage 1 (Release 17). 3GPP, July 2020
[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 V15.7.0: System Architecture for the 5G System (5GS); Stage 2 (Release 15). 3GPP, September 2019
[38] TS 23.501 V16.2.0: System Architecture for the 5G System (5GS); Stage 2 (Release 16). September 2019
[39] TS 23.502 V15.7.0: Procedures for the 5G System (5GS); Stage 2 (Release 15). 3GPP, September 2019
[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 V15.3.0: Management and Orchestration; Concepts, use cases and requirements (Release 15). 3GPP, January 2020
[43] TS 29.500 V15.5.0: 5G System; Technical Realization of Service Based Architecture; Stage 3 (Release 15). 3GPP, September 2019
[44] TS 29.501 V15.6.0: 5G System; Principles and Guidelines for Services Definition; Stage 3 (Release 15). 3GPP, December 2019
[45] TS 33.501 V15.7.0: Security architecture and procedures for 5G system (Release 15). 3GPP, December 2019
[46] TS 37.340: 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 V15.7.0: NR; NR and NG-RAN Overall Description; Stage 2 (Release 15). 3GPP, September 2019
[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/specifications-groups/specifications-groups
[56] https://www.3gpp.org/contact/3gpp-faqs
[57] https://www.3gpp.org/specifications/67-releases
[58] https://www.3gpp.org/3gpp-calendar/44-specifications/releases
[59] https://www.3gpp.org/release-15
[60] https://www.3gpp.org/release-16
[61] https://www.3gpp.org/release-17
[62] https://www.3gpp.org/news-events/2098-5g-in-release-17-%E2%80%93-strong-radio-evolution
[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] https://www.bfs.de/DE/themen/emf/kompetenzzentrum/mobilfunk/basiswissen/5g.html
[80] https://www.bfs.de/DE/themen/emf/kompetenzzentrum/mobilfunk/schutz/grenzwerte.html
[81] https://www.bundesnetzagentur.de
[82] https://www.bundesnetzagentur.de/DE/Sachgebiete/Telekommunikation/Unternehmen_Institutionen/Breitband/MobilesBreitband/Frequenzauktion/2019/
[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, Entwurf. Bundesnetzagentur, 9.10.2019
[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; Röösli, 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, 2018

[94] http://www.etsi.org/technologies-clusters/technologies/multi-access-edge-computing

[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 White Paper, 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 – White Paper. 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, 2013

[119] Karaboytcheva, Miroslava: Auswirkungen der drahtlosen 5G Kommunikation auf die menschliche Gesundheit. Briefing. EPRS – Wissenschaftlicher Dienst des Europäischen Parlaments, PE 646.172, Februar 2020

[120] Keith, Robert: 5G Energy Efficieny Explained. A10 Networks, 10.3.2020 (https://www.a10networks.com/blog/5g-energy-efficiency-explained)

[121] Kühn, Paul: Vorlesungsskript Nachrichtenvermittlung I und II. Universität Stuttgart, Institut für Nachrichtenvermittlung und Datenverarbeitung, 1991

[122] Latva-aho, Matti; Leppänen, Kari: Key Drivers and Research Challenges for 6G Ubiquitous Wireless Intelligence. 6G Research Visions 1. 6G Flagship, University of Oulu, September 2019

[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 (https://de.slideshare. net/veermalik121/5g-new-radio-technology-introduction-and-its-throughput-capabilities)

[132] GS NFV-MAN 001 V1.1.1: Network Functions Virtualisation (NFV); Management and Orchestration. ETSI, December 2014

[133] Marsch, Patrick; Bulakçi, 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] NTT: 5G Evolution and 6G, White Paper. NTT DOCOMO, January 2020

[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] Peisa, Janne; Persson, Patrik; Parkvall, Stefan; Dahlman, Erik; Grovlen, Asbjorn; Hoymann, Christian; Gerstenberger, Dirk: 5G evolution: 3GPP Releases 16 & 17 Overview. Ericsson Technology Review, March 9, 2020

[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 8 – Up to June 2020, Study for European Commission. IDATE DIGIWORLD, July 2020

[156] Rijsman, B.; Moisand, J.: draft-rijsman-sfc-metadata-considerations-00.txt – Metadata Considerations. IETF, February 2014

[157] https://www.rtr.at

[158] Sauter, Martin: From GSM to LTE-Advanced Pro and 5G. Wiley, 2017

[159] 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, 2014

[162] Siegmund, Gerd: Technik der Netze 2. VDE, 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: White Paper – 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 – White Paper. 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] Zhang, Zhengquan; Xiao, Yue; Ma, Zheng; Xiao, Ming; Ding, Zhiguo; Lei, Xianfu; Karagiannidis, George K.; Fan, Pingzhi: 6G Wireless Networks – Vision, Requirements, Architecture, and Key Technologies. IEEE Vehicular Technology Magazine, vol. 14, no. 3, pp. 28-41, September 2019

[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] Delb, Valentin; Dudle, Gregor; Dürrenberger; Gregor; Grasser, Christian; Horisberger, Philippe; Künzle, Harry; Kuster, Niels; Netzle, Stephan; Portmann, Manfred; Quinto, Carlos; Reichenbach, Alexander; Röösli, 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

[186] http://www.3gpp2.org

[187] Lescuyer, Pierre; Lucidarme, Thierry: EvolvedPacket System (EPS) – The LTE and SAE Evolution of 3G UMTS. Wiley, 2008

[188] https://www.5gamericas.org

[189] http://www.imt-2020.cn/en

[190] http://www.5gforum.org/html/en/main.php

[191] https://5gmf.jp/en

[192] https://tiaonline.org/website_categories/5g/

[193] https://www.fcc.gov/auctions

[194] https://www.fcc.gov/

[195] https://www.fcc.gov/5G

[196] 5G Americas: 5G Spectrum Vision. Whitepaper. 5G Americas, February 2019

[197] https://5gobservatory.eu/5g-spectrum/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] https://www.cooley.com/news/insight/2020/2020-03-11-fcc-sets-auction-schedule-for-2020

[201] 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

[202] FG NET-2030 Sub-G1 Technical Report: Representative Use Cases and Key Network Requirements for Network 2030. ITU-T FG NET-2030, January 2020

[203] 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

[204] FG NET-2030 Sub-G2 Technical Specification: Network 2030 Architecture Framework. ITU-T FG NET-2030, June 2020

[205] 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

Index

https://doi.org/10.1515/9783110724509-015