1

Introduction

This book is about *fifth-generation* (5G) mobile telecommunications. Our first chapter puts 5G into its historical context, and lays out its requirements and its most important technical features. We begin by setting out the architecture of a mobile telecommunication system and by reviewing the history of mobile telecommunications. We then discuss the underlying technologies that 5G uses, review the standardization process, and introduce the capabilities and architecture of the 5G system. The chapter closes with a survey of the use cases and applications for 5G, and a review of the current state of the market.

1.1 Architecture of a Mobile Telecommunication System

1.1.1 High-level Architecture

A mobile telecommunication system is run by a *network operator*, such as Vodafone, AT&T or China Mobile, and is officially known as a *public land mobile network* (PLMN). As shown in Figure 1.1, it has four main components, namely the core network (CN), the radio access network (RAN), the management system and the user's device. The last of these is known colloquially as the *mobile*, and more formally as the *user equipment* (UE).

The core network transports traffic between the mobile and one or more external networks, such as the *public switched telephone network* (PSTN) or the internet. The core network also controls the mobile's communications with those external networks, and stores information about the network operator's subscribers.

The radio access network handles the network's radio communications with the mobile. It communicates with the core network over an interface known as the *backhaul*, and with the mobile over the *air interface*, also known as the radio interface. On that interface, the direction from network to mobile is known as the *downlink* (DL) or forward link, and the direction from mobile to network is the *uplink* (UL) or reverse link.

The network is controlled by a separate *management system*. Its tasks include configuring the various components of the core and radio access networks, monitoring their performance, reporting any faults to the network operator and billing the user.

A mobile can communicate outside the coverage area of its network operator by using the resources from two PLMNs, namely the *visited network*, where the mobile is located, and the operator's *home network*. That situation is known as *roaming*.

An Introduction to 5G: The New Radio, 5G Network, 5G Advanced and Beyond, Second Edition. Christopher Cox.

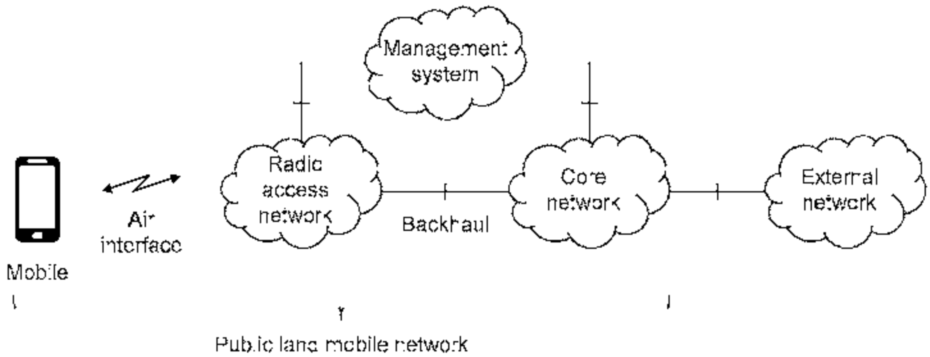

Figure 1.1 Architecture of a mobile telecommunication system.

1.1.2 Internal Architecture of the Mobile

Internally, the mobile has two main components. The actual communication device is known as the *mobile equipment* (ME). In the case of a voice mobile or a smartphone, that is just a single device. However, the ME can itself have two parts, namely the *mobile termination* (MT), which handles all the communication functions, and the *terminal equipment* (TE), which terminates the data streams. The MT might be a plug-in 5G card for a laptop, for example, in which case the TE would be the laptop itself.

The *universal integrated circuit card* (UICC) is a smart card, colloquially known as the SIM card. It runs an application known as the *universal subscriber identity module* (USIM), which stores user-specific data such as the home network identity and carries out security-related calculations using secure keys that the smart card stores.

1.1.3 Architecture of the Radio Access Network

The most important component of the radio access network is the *base station*. In a medium-sized country like the United Kingdom, a typical network might contain several thousand of these. Figure 1.2 shows a group of nearby base stations, as viewed from above.

On the air interface, a base station transmits and receives using one or more radio frequencies, each of which is known as a *carrier frequency*. Around each carrier, the radio signal occupies a certain amount of radio spectrum, which is known as the *bandwidth*. For example, a 5G base station might use a carrier frequency of 3500 MHz and a bandwidth of 40 MHz, so that its transmissions span the range from 3480 to 3520 MHz.

Usually, the base station operates in *licensed spectrum*, in which the network operator has purchased an exclusive licence to carry out radio communications from the corresponding national regulator. The alternative is *unlicensed spectrum*, in which there are no licences, but the transmitters have to use a low power so that they do not cause undue interference to other receivers in the same band.

The base station controls one or more *cells*, each of which is a radio transmission with a particular carrier frequency and bandwidth, which spans a particular coverage area. On any one radio frequency, a base station can control multiple cells, known as *sectors*, by transmitting in different directions. There are typically three sectors per base station, with each one spanning

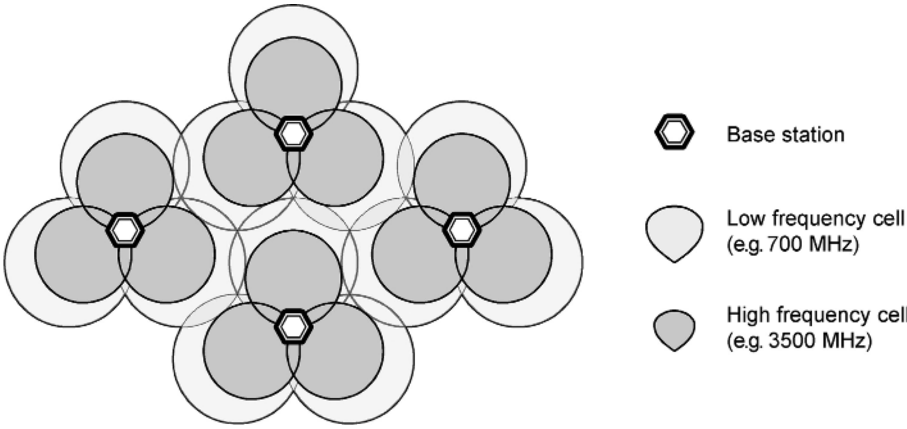

Figure 1.2 Example of base stations with two carrier frequencies and three sectors.

an arc of 120°. The base station can also control multiple cells in the same direction by using different radio frequencies to ensure that the radio signals do not interfere.

1.1.4 Coverage and Capacity

There are two main limits on a cell's performance. Firstly, each cell has a coverage limit, which is the maximum distance at which the receiver can successfully hear the transmitter. A cell's coverage is greatest if the transmissions are at a high power and if the data rate is low because both conditions imply that the energy per bit at the receiver is high. In a simple communication system, in which the base station and mobile both have a single antenna, the cell's coverage is also greatest if the carrier frequency is low, for example 700 MHz rather than 3500 MHz.

Secondly, each cell has a capacity limit, which is the maximum combined data rate of all the mobiles that are communicating through the cell. A cell's capacity is greatest if the transmissions occupy a wide bandwidth and are received with a high *signal-to-interference plus noise ratio* (SINR), which is the power at the receiver from the desired signal divided by the power from noise and interference. More precisely, a cell's capacity is constrained by the *Shannon limit* [1], which in a single-antenna communication system is as follows:

$$C = B \log_2(1 + \text{SINR}) \tag{1.1}$$

where B is the bandwidth in Hz and C is the cell's capacity in bits s^{-1}. Improvements in radio communication technology allow the cell's capacity to approach the Shannon limit, but not to exceed it, while a more general version of the equation allows a higher capacity in a multiple-antenna system [2].

Bearing those limits in mind, cells can be grouped into a number of different classes. *Macrocells* are a few kilometres across and provide wide-area coverage in rural areas and suburbs, with the signals travelling over the building roofs. *Microcells* are a few hundred metres across and are used in urban areas, with the signals confined to individual streets by the buildings that surround them. *Picocells* are a few tens of metres across and are used for indoor communications in shopping centres and offices. *Femtocells* are a few metres across and are installed by consumers within the home. The same area can be spanned by one macrocell or by a large number of smaller cells, so the collective capacity of those small cells can be very large.

1.1.5 Architecture of the Core Network

In older systems, the core network contains two domains, which transport different types of traffic using different networking techniques. The *circuit-switched* (CS) domain transports fixed-rate traffic such as voice, so that the user can make phone calls with other devices that are in the public switched telephone network or the circuit-switched domains of other network operators. It does that using a technique known as *circuit switching*, which sets aside a dedicated two-way connection for each individual phone call. The CS domain transports voice traffic with a constant data rate and minimal delay, but it is inappropriate for services in which the data rate can vary.

The *packet-switched* (PS) domain transports variable-rate data traffic, such as web pages and emails, between the user and external data networks such as the internet. It does that using a different technique, known as *packet switching*, in which a data stream is divided into packets, each of which is labelled with the address of the required destination device. Within the network, *routers* read the destination addresses of the incoming data packets and forward them towards those destinations by following the instructions in internal *routing tables*. The network's resources are shared amongst all the users so the technique is more efficient than circuit switching. However, delays can result if too many devices try to transmit at the same time.

In recent years, the use of smartphones has led to a situation in which mobile network traffic is dominated by data rather than voice. In response, designers have abandoned the use of circuit switching and have introduced mobile communication systems that use packet switching alone. That simplifies the design and allows the system to be optimized for the delivery of data traffic, but it implies that voice calls have to be handled in other ways. We will discuss those in Section 1.2.4.

1.1.6 Communication Protocols

In any communication network, information is transported using hardware and software functions known as *protocols*. The best examples are the protocols used for packet-switched data transport over the internet, because those protocols are also used by the core and radio access networks of 5G. The internet's protocols are designed by the *Internet Engineering Task Force* (IETF) and are grouped into various numbered *layers*, each of which handles one aspect of the transmission and reception process. The usual numbering follows a seven-layer model known as the *Open Systems Interconnection* (OSI) model, with the top three OSI layers folded together in the case of the internet.

As an example (Figure 1.3), let us suppose that a web server is sending information to a user's browser. In the first step, an *application layer* protocol – in this case, the *Hypertext Transfer Protocol* (HTTP) – receives information from the server's application software and passes it to the next layers down by representing it in a way that the user's application layer will eventually be able to understand.

The *transport layer* manages the end-to-end data transmission. There are two main alternatives. The *Transmission Control Protocol* (TCP) re-transmits a packet from end to end if it does not arrive correctly and is suitable for data such as web pages and emails that have to be received reliably. The *User Datagram Protocol* (UDP) sends the packet without any re-transmission and is suitable for information such as real-time voice or video for which timely arrival is more important.

In the *network layer*, the *Internet Protocol* (IP) sends packets on the correct route from source to destination, using the *IP address* of the destination device. The process is handled by the

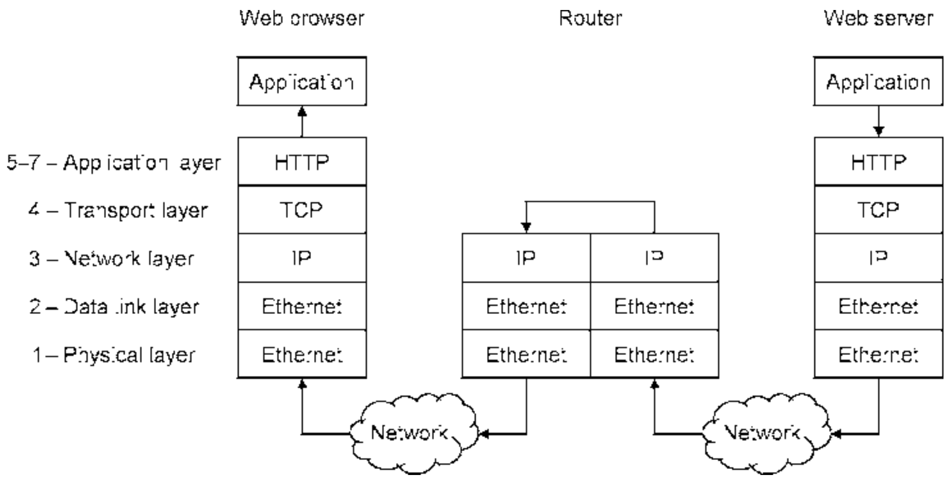

Figure 1.3 Examples of the communication protocols used by the internet.

intervening routers, which inspect the destination IP addresses by implementing just the lowest three layers of the protocol stack. The *data link layer* manages the transmission of packets from one device to the next, for example by re-transmitting a packet across a single interface if it does not arrive correctly. Finally, the *physical layer* deals with the actual transmission details, for example by setting the voltage of the transmitted signal. The internet can use any suitable protocols for the data link and physical layers, such as *Ethernet*.

At each level of the transmitter's stack, a protocol receives a data packet from the protocol above in the form of a *service data unit* (SDU). It processes the packet, adds a header to describe the processing it has carried out and outputs the result as a *protocol data unit* (PDU). This immediately becomes the incoming SDU of the next protocol down. The process continues until the packet reaches the bottom of the protocol stack, at which point it is transmitted. The receiver reverses the process, using the headers to help it undo the effect of the transmitter's processing.

1.2 History of Mobile Telecommunications

1.2.1 Introduction

In the time since they were first introduced in the early 1980s, a new *generation* of mobile communication systems has appeared every 10 years or so. The term generation emerges by industry consensus: it does not have any formal definition, but it is typically associated with a new air interface and with a new architecture for the network.

The *first-generation* (1G) systems used analogue communication techniques that were similar to those in a traditional analogue radio. The individual cells were large and the systems did not use the available radio spectrum efficiently, so their capacity was by today's standards very small. The mobile devices were large and expensive, and were marketed almost exclusively to business users. Some popular systems included *Nordic Mobile Telephone* (NMT) in Scandinavia, the *Advanced Mobile Phone System* (AMPS) in the USA and the *Total Access Communication System* (TACS) in the UK.

1.2.2 Global System for Mobile Communications (GSM)

Mobile telecommunications took off as a consumer product with the introduction of *second-generation* (2G) systems in the early 1990s. Those systems were the first to use digital technology, which permitted a more efficient use of radio spectrum and the introduction of smaller, cheaper devices. The most popular 2G system was the *Global System for Mobile Communications* (GSM), which was originally specified by the *European Telecommunications Standards Institute* (ETSI).

GSM was originally designed just for voice, but was later enhanced to support instant messaging by means of the *Short Message Service* (SMS), and the delivery of packet data to a mobile device by means of the *General Packet Radio Service* (GPRS). Another enhancement was *Enhanced Data Rates for GSM Evolution* (EDGE), which modified the air interface to increase the available data rate.

In GSM, the core network has the traditional two domains: the circuit-switched domain handles voice calls, while the packet-switched domain handles data. The RAN is known as the *GSM EDGE radio access network* (GERAN), while the base station is known as the *base transceiver station* (BTS). Figure 1.4 shows the high-level architecture of GSM and its descendants, and highlights the most important interfaces.

1.2.3 Universal Mobile Telecommunication System (UMTS)

Buoyed by the success of 2G, the *International Telecommunication Union* (ITU) published a set of requirements in 1997 for a *third-generation* (3G) mobile communication system, under the name *International Mobile Telecommunications* (IMT) *2000* (IMT-2000) [3]. A number of standards bodies responded by submitting specifications, which the ITU eventually accepted as meeting the requirements of IMT-2000. The most popular was the *Universal Mobile Telecommunication System* (UMTS), which was designed by the *Third Generation Partnership Project* (3GPP) [4], a specially formed collaboration that later took responsibility for the specifications for GSM.

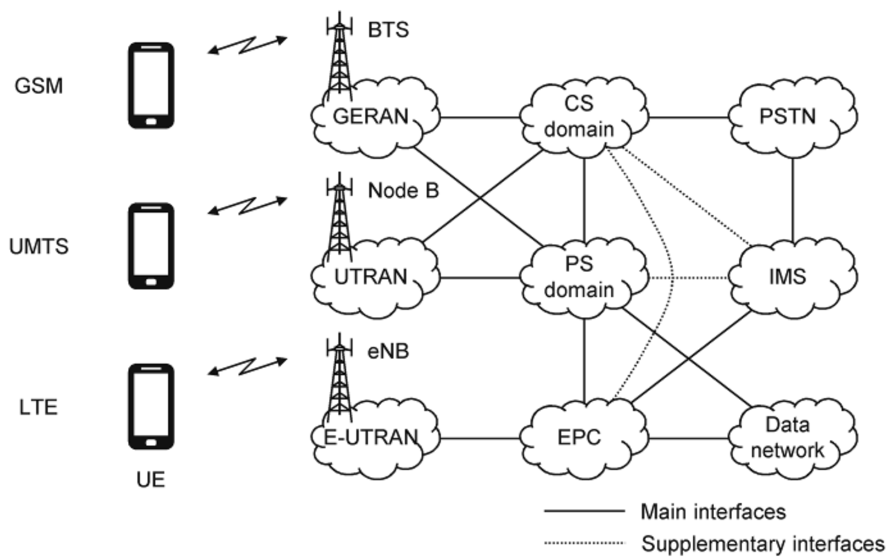

Figure 1.4 High-level architectures of GSM, UMTS and LTE.

UMTS uses the same core network as GSM, but has a new RAN known as the *UMTS terrestrial radio access network* (UTRAN) and a new base station known as the *Node B*. The UMTS air interface is more efficient than GSM's, and has two implementations: *wideband code division multiple access* (WCDMA) is used throughout most of the world, while *time division synchronous code division multiple access* (TD-SCDMA) is derived from WCDMA and is mainly used in China. The most important later enhancement is known as *high-speed packet access* (HSPA), which improves the performance of data applications by increasing the air interface's average data rate, at the expense of short-term variations in the data rate for an individual user.

1.2.4 Long-term Evolution (LTE)

The introduction of HSPA coincided with the introduction of smartphones, notably the Apple iPhone in 2007, followed by devices based on Google's Android operating system from 2008. Together, those factors led to an explosion in the use of mobile data applications and in the amount of mobile data traffic.

In response, *fourth-generation* (4G) systems are optimized for the delivery of data in the form of *mobile broadband*. By far the most popular is the 3GPP system known as *Long-term Evolution* (LTE). In LTE, the base station is known as an *evolved Node B* (eNB), the RAN is the *evolved UMTS terrestrial radio access network* (E-UTRAN) and the *evolved packet core* (EPC) replaces the packet-switched domain of UMTS and GSM. Officially, the whole system is known as the *evolved packet system* (EPS), although in practice that term is used only occasionally.

LTE does not have a circuit-switched domain of its own, but there are still various ways for an LTE user to make a voice call. The first one specified was *circuit-switched fallback* (CSFB), in which the network transfers the mobile to a legacy 2G or 3G cell, through which the mobile contacts the CS domain of the 2G or 3G core. Alternatively, the mobile can make a *voice over IP* (VoIP) call by means of the EPC, using a third-party application such as Skype or WhatsApp. Finally, network operators can provide their own IP voice services by means of an external 3GPP network known as the *IP multimedia subsystem* (IMS), in a technology known as *voice over LTE* (VoLTE). Although rather complex, the IMS brings several advantages, for example it can handle services such as emergency calls that are not supported by third-party applications, it maintains the operators' voice revenue and it frees network operators from reliance on nearby 2G or 3G cells. It also supports a technique known as *single radio voice call continuity* (SRVCC), in which the network can hand a mobile over to a 2G or 3G cell if its leaves the coverage area of LTE, and can convert an ongoing VoIP call to a circuit-switched call that is handled by the 2G or 3G core.

1.2.5 LTE-Advanced

In 2008, the ITU published requirements for a successor to IMT-2000, which is known as *IMT-Advanced* [5]. The requirements called for peak downlink and uplink data rates of 1000 and 500 Mbps, respectively, which were beyond the original capabilities of LTE. 3GPP therefore responded by writing specifications for an enhanced version of LTE, known as *LTE-Advanced*.

The most important enhancement in LTE-Advanced is known as *carrier aggregation* (CA). This technique increases a mobile's data rate by allowing it to communicate through multiple cells, which are transmitted and received on different radio frequencies, but are still controlled by a single eNB.

A later enhancement, known as *dual connectivity* (DC), allows those cells to be controlled by two separate base stations rather than by one. The original base station carries traffic for the user and signalling messages that control the mobile, and is known as the *master eNB* (MeNB). Typically, the MeNB controls one or more macrocells, which provide the mobile with wide-area coverage and reliable communications. The additional base station mainly carries traffic and is known as the *secondary eNB* (SeNB). Typically, the SeNB controls one or more microcells, which provide mobiles that can receive them with a much higher data rate. Dual connectivity was not widely implemented at first, but it was an important step towards the introduction of 5G.

1.2.6 LTE-Advanced Pro

Later enhancements to LTE are known by the name *LTE-Advanced Pro*. The enhancements broaden the capabilities that LTE offers and the services that it supports, and act as a precursor to 5G.

LTE was originally designed for use by people, but there is also a need for wide-area *machine-type communications* (MTC) for devices such as vehicle trackers and smart utility meters. LTE supports those devices using two enhancements. *Enhanced machine-type communications* (eMTC), also known as *LTE machine type communication* (LTE-M), is backwards compatible with LTE, in the sense that an eMTC base station can communicate with both legacy mobiles and machine-type devices at the same time. (On the other hand, a legacy base station cannot communicate with a machine-type device that only supports eMTC.) The *narrowband internet of things* (NB-IoT) has better performance characteristics, but is defined as a separate air interface technology that is not backwards compatible with LTE: an NB-IoT base station cannot communicate with legacy mobiles, and vice versa.

The LTE specifications have always allowed a mobile to access the evolved packet core by means of a non-3GPP access network. The most common choice is a *wireless local area network* (WLAN), also known as WiFi, which is specified by the *Institute of Electrical and Electronics Engineers* (IEEE) in a set of standards denoted IEEE 802.11. Later enhancements build on that capability in various ways. *Licence-assisted access* (LAA) is based on carrier aggregation, but it introduces additional LTE cells into the unlicensed spectrum that WiFi would normally use. *LTE WLAN aggregation* (LWA) is based on dual connectivity, except that the mobile communicates with a master eNB and a WiFi access point, rather than a master and secondary eNB.

Device-to-device (D2D) communications allow two mobiles to communicate directly with one another over a new air interface that is known as the *sidelink* (SL). The initial application is in public safety networks such as those used by the police and other emergency services, in which such communication is important if the base station has been taken down by an emergency situation or is otherwise out of range. D2D communication is also a useful way to offload traffic from the base station, for applications such as file exchange between nearby devices.

The sidelink is also a valuable part of vehicle-related communications, in other words the exchange of information to and from a road vehicle. Although full support for vehicle communications requires 5G, LTE can address some of the less stringent, non-safety-critical aspects. Enhancements include support for ETSI's architecture for an *intelligent transport system* (ITS), and improvements to the sidelink so that vehicles can communicate even if their relative speed is high.

1.2.7 Other Mobile Communication Systems

Although 3GPP systems have come to dominate the mobile communication market, two other technology families have been important. *IS-95* was a technology designed by Qualcomm, which became the dominant 2G system in the United States. It was succeeded by a 3G system known as *cdma2000*, which was specified by a similar collaboration to 3GPP known as the *Third Generation Partnership Project 2* (3GPP2), and which met the requirements for IMT-2000. Qualcomm intended to develop a 4G successor to cdma2000 under the name *Ultra Mobile Broadband* (UMB), but no network operator announced plans to adopt the technology and the project was dropped in 2008.

Worldwide Interoperability for Microwave Access (WiMAX) is specified in the IEEE 802.16 standards. The original specification was for a point-to-point microwave system, but that gradually evolved into *Mobile WiMAX 1.0* (IEEE 802.16e), which had all the characteristics of a mobile cellular network and met the requirements for IMT-2000. The IEEE later specified an enhancement known as *Mobile WiMAX 2.0* (IEEE 802.16m), which met the requirements of IMT-Advanced. However, LTE had far more support amongst network operators and equipment manufacturers, and WiMAX no longer has a significant market share.

1.2.8 Early Research into 5G

In 2012, the ITU started a programme of work to define a successor to IMT-Advanced, which it denoted *IMT-2020* [6–10]. That triggered a surge of national and industrial research projects into 5G, which are summarized in Reference [11]. It also triggered a 3GPP study item that began in 2015 [12].

These research projects identified several potential markets and a large number of individual use cases for 5G, and grouped those use cases into three families that share common requirements and characteristics. Two of them build on the capabilities of LTE: *enhanced mobile broadband* (eMBB) addresses the delivery of mobile data to individual consumers [13], while *massive machine-type communications* (mMTC) refers to communications with large numbers of autonomous machine-type devices [14]. *Ultra-reliable low-latency communication* (URLLC) is new to 5G and addresses applications that require very reliable communications in conjunction with a very low delay [15]. We will discuss these use case families in Sections 1.5 to 1.7, along with some more specialized applications that have been addressed as 5G has evolved.

The diverse nature of these use cases implies that 5G has to support a broader set of requirements than previous systems have done. To help it do so, 3GPP defined a further set of internal requirements for *network operation* (NEO) [16]. These include, as examples, requirements to optimize and re-route the traffic path in the case of low-latency applications, to interact with a client's application servers and to reconfigure a network to support new use cases as they arise. Those requirements are addressed by a number of enabling technologies for the 5G system, which we will now introduce.

1.3 Technologies for 5G

1.3.1 Network Function Virtualization

The core network of a mobile telecommunication system contains a large number of network elements, for example routers, signalling functions and databases. Traditionally, these have been implemented using special-purpose hardware, which is procured from an equipment vendor such as Ericsson, Nokia or Huawei. That approach is perfectly satisfactory, but has several limitations: the hardware is manufactured in low volumes, so it is expensive; the hardware and software are coupled together, with the operator locked to a single vendor for both; and it is difficult, slow and expensive to upgrade the hardware as new requirements arise, an issue that becomes ever more problematic as development lifecycles shorten.

Network function virtualization (NFV) is a more recent approach in which the network elements are implemented using software in the form of *virtualized network functions* (VNFs) [17–19]. Those functions then run on general-purpose, commercial off-the-shelf hardware. Network operators were already adopting these techniques in their existing networks [20]. However, 5G has been designed with NFV in mind from the beginning, for example in the design of a management system which can easily set up, modify and tear down the individual network functions, and which can administer them independently of the underlying hardware [21].

The use of NFV brings several benefits, for example the hardware is generic and cheaper, while the software is easier, faster and cheaper to write and upgrade. NFV also facilitates a concept known as *disaggregation*, in which the software is broken down into small, clearly defined components. Those components are separated from the hardware and from each other using *open interfaces*, which are fully standardized in the interests of interoperability. In a disaggregated network, the small component size makes it easier for new suppliers to enter the market. In turn, the use of open interfaces allows operators to obtain their network functions from multiple suppliers, so that they are no longer locked to a single one.

1.3.2 Software-defined Networking

At the highest level, each element in a traditional communication network handles two main tasks. Functions in the *user plane* (UP) forward traffic from one network element to another, for example by the use of routing tables, while those in the *control plane* (CP) carry out higher-level tasks such as configuring the routing tables and managing the network's resources. Once again, that approach is perfectly satisfactory, but has several limitations: the decision-making processes are distributed over the entire network, which makes the control plane unnecessarily complex, while those complexities limit how far the operator can distribute the functions in the user plane.

The central feature of *software-defined networking* (SDN) is a clear separation between the control and user planes [22]. In a software-defined network, the control plane functions can be centralized, which allows for simpler, more integrated control over the network. Once centralized, those functions can easily communicate with authorized third-party application servers over a so-called *northbound application programming interface* (API), through which those servers can influence or control the network's operation. The third-party servers do not know any details of the network's physical implementation, as those are confined to the *southbound API*, through which the control plane functions configure the ones in the user plane. In turn, the user plane functions can be distributed to whichever physical locations are desired, without the constraints that the control plane previously imposed.

NFV and SDN are complementary technologies: the first separates the hardware and software in a communication network, while the second separates its control and user planes. Once again, 5G has been designed with SDN in mind from the beginning.

1.3.3 Network Slicing

NFV and SDN provide a network operator with a flexible underlying resource, which is easily reconfigured to handle changes in network traffic. Taking those concepts further, *network slices* are virtual logical sub-networks which run on shared, underlying physical hardware, and which are realized by means of NFV and SDN [23, 24]. A network operator can set up, reconfigure and tear down individual slices as the need for them arises.

Each network slice is optimized for the traffic that it carries. For example, a network operator might use different slices for each of the main 5G use cases, with the eMBB, mMTC and URLLC slices, respectively, optimized for a high data rate, a high connection density, and low latency and high reliability. An operator might use different slices to subdivide those use cases, for example the eMBB use case might be realized using different slices for internet traffic, IMS voice and video, and public safety. Different slices might also be used for an operator's individual customers, such as *mobile virtual network operators* (MVNOs) that do not own any infrastructure of their own, or industrial clients.

An operator's network slices are isolated from each other, most critically in respect of data transport and security. If, for example, there is congestion in a slice that is handling internet traffic, then there should be no additional delays in slices that are handling time-sensitive applications such as IMS voice, public safety or URLLC. Similarly, any intrusion into a slice that is handling MTC should not increase the risk of security issues in any of the other network slices.

LTE networks already implement some of the ideas underlying network slicing, for example LTE transport networks often isolate IMS voice from internet data to minimize congestion problems. However, network slicing can only be fully realized in a 5G network by means of NFV and SDN.

1.3.4 Technologies for the Radio Access Network

The radio access network can also make use of these technologies, but the terminology can be confusing. A traditional LTE network distributes all of the base stations' components across the local cell sites and is known as a *distributed RAN* (D-RAN). In contrast, a *centralized RAN* (C-RAN) implements the higher-level functions of the base station at a more central hub, which can be many kilometres from the cell site, but should still be close enough for low-latency communication. By itself, RAN centralization offers improved security and a smaller footprint at the cell site, and allows the higher-level functions of nearby cells to communicate more efficiently.

Moving to network function virtualization, a *virtualized RAN* (V-RAN) implements most of the base station in software on commercial off-the-shelf hardware, with hardware accelerators reserved for only a few of the lowest-level functions. If the concepts of virtualization and centralization are used together then the result is a *cloud RAN* (also C-RAN), in which higher-level parts of the base station are implemented in a pool of centralized, off-the-shelf servers. As well as bringing the generic benefits of NFV, a cloud RAN allows for load balancing, in which heavily loaded base stations can acquire more resources from the underlying hardware than their lightly loaded neighbours.

An *open RAN* breaks down the base station into smaller software functions by means of disaggregation and virtualization, and separates those functions by means of open, interoperable interfaces. That eases the entry of new vendors into the market and allows an operator to mix components that have been procured from different vendors. The name *open RAN* is sometimes abbreviated as O-RAN, but that abbreviation is usually reserved for a specific set of specifications that have been developed by an industry body known as the *O-RAN Alliance* [25]. We will discuss those specifications as part of Chapter 3.

Software defined networking can be applied to the radio access network as well. *Multi-access edge computing* (MEC), also known as *mobile edge computing*, is a technology in which the user plane connection to an external network is implemented close to the cell site, or even at the cell site itself [26–29]. By shortening the communication path, the architecture provides the mobile

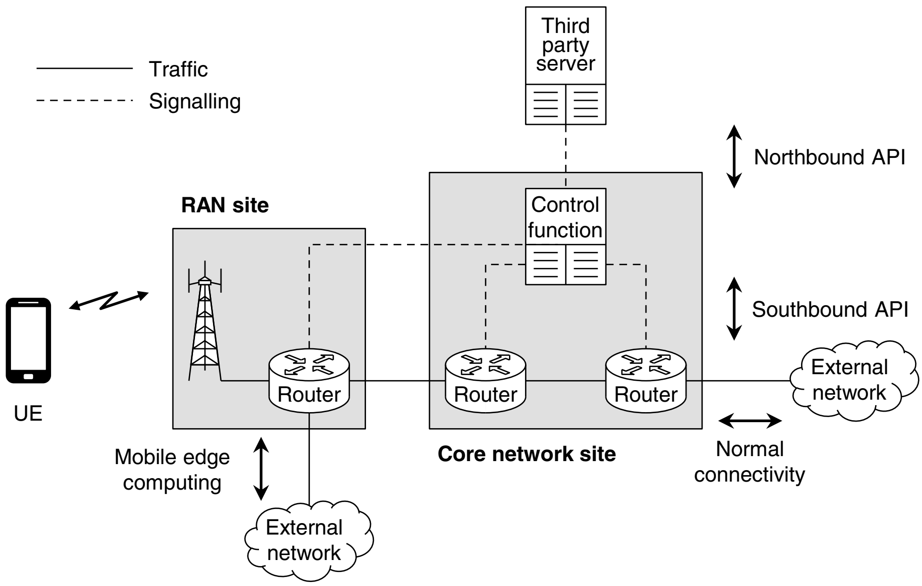

Figure 1.5 Architecture of a telecommunication network using software-defined networking and mobile edge computing.

with low-latency access to external servers to help support the most demanding forms of low-latency application. Furthermore, if the user plane function is implemented in software by means of NFV, then it can be relocated dynamically, for example if the mobile starts to use a low-latency service or if it subsequently changes cell. Figure 1.5 is a high-level view of the resulting architecture for a mobile telecommunication network that is using the two technologies.

With the capabilities of NFV and SDN in place, a 5G network slice extends not only through the core network, but also into the radio access network and the air interface itself. In that context, the air interface can be understood as another resource that is easily shared amongst network slices as the need arises.

1.3.5 Technologies for the Air Interface

The 5G air interface uses two important technologies that we will address in detail in later chapters. Firstly, 5G can operate over a wider range of radio frequencies than previous systems have done, up to a maximum of about 70 GHz. At the highest frequencies the radio waves have a wavelength of a few millimetres, so they are often known as *millimetre waves*. When using millimetre waves, the 5G air interface can occupy a far wider bandwidth than it does at low radio frequencies, so its capacity can be far greater. However, the use of millimetre waves also leads to a number of coverage issues, which we will explore as part of Chapter 4.

Secondly, 5G makes greater use of multiple antennas than previous systems have done [30], typically using just a few antennas at the mobile, but using tens or even a few hundred at the base station. It applies them in two distinct ways: to improve the coverage of the air interface at high radio frequencies by means of *beamforming*, and to improve its capacity at low radio frequencies by means of *multiple-input multiple-output* (MIMO) antennas. We will examine both of these techniques as part of Chapter 6.

1.3.6 Artificial Intelligence and Machine Learning

Alongside these technologies, 5G has been making increased use of *artificial intelligence and machine learning* (AI/ML). The two terms overlap, but they are distinct [31]. *Artificial intelligence* (AI) is the ability of a computer to mimic aspects of human intelligence. The underlying task is usually one of *inference*, in which an AI model takes a complex input and produces a simple output, for example by classifying a picture as a cat or a dog. That in turn supports more complex tasks, for example by creating a new picture based on simple text prompts.

The AI model is usually implemented by an *artificial neural network*. That mimics the operation of the brain by combining the inputs in a calculation that is conceptually simple but uses a large number of adjustable parameters. The details are usually proprietary: the inputs and outputs are clearly defined, but the calculations are hidden.

The parameters' values are not explicitly defined. Instead, they are computed during a process of *machine learning* (ML), also known as *training*. During that process, the AI/ML model takes a set of *training data* and calculates the corresponding output. The model's parameters are then adjusted to bring the actual output slightly closer to the desired one. After many such steps involving large amounts of training data, the parameters have been optimized and the model is said to be trained.

There are a few different implementations. During supervised machine learning, the desired output is identified by a human operator, for example by labelling a training image as a cat or a dog. Unsupervised learning involves the automatic identification of features that distinguish the input data, while semi-supervised learning uses a mix of the two. During reinforcement learning, the model has a simple goal (e.g. winning a game of chess) and learns the best way to achieve that goal by trial and error.

Later releases of 5G include a framework for the use of artificial intelligence and machine learning throughout the 5G system [32]. Training and inference are handled either by the management system or by the individual network functions, while the signalling procedures are enhanced to support the deployment of AI models and the exchange of training and inference data. We will see some examples as part of Chapters 20 and 23.

1.4 The 3GPP 5G System

1.4.1 The 3GPP Specifications for 5G

The 5G specifications are written by 3GPP in the same way as the earlier specifications for LTE, UMTS and GSM. They are organized into *releases* [33], each of which contains a stable and clearly defined set of features. Within each release, the specifications progress through a number of different versions. New features can be added to successive versions until the release is frozen, after which the only changes involve refinement of the technical details, corrections and clarifications. Table 1.1 lists the releases that 3GPP has used since the introduction of UMTS, together with their most important features. (The numbering scheme was changed after Release 99, with later releases numbered from 4 onwards.)

5G was first specified in Release 15, which focuses mainly on eMBB, but also includes support for low-latency and high-reliability communications. Within that release, the specifications were frozen in three separate drops to help equipment manufacturers introduce the most important capabilities early on without waiting for later capabilities to be specified in full [34]. Release

Table 1.1 3GPP specification releases for UMTS, LTE and 5G.

Release	End date	New features
Rel-99	December 1999	UMTS, WCDMA
Rel-4	June 2001	TD-SCDMA
Rel-5	September 2002	HSDPA, IP multimedia subsystem
Rel-6	September 2005	HSUPA
Rel-7	March 2008	Enhancements to HSPA
Rel-8	March 2009	LTE
Rel-9	March 2010	Location services for LTE
Rel-10	June 2011	LTE-Advanced, carrier aggregation
Rel-11	March 2013	Co-ordinated multi-point transmission
Rel-12	March 2015	Dual connectivity, LTE sidelink
Rel-13	March 2016	LTE-Advanced Pro, eMTC, NB-IoT
Rel-14	June 2017	Vehicle communications for LTE
Rel-15	June 2019	5G
Rel-16	June 2020	MTC and vehicle communications for 5G
Rel-17	June 2022	Non-terrestrial networks
Rel-18	June 2024	5G-Advanced, extended reality, AI/ML
Rel-19	December 2025 (target)	Enhancements to 5G-Advanced
Rel-20	June 2027 (target)	Studies into 6G
Rel-21	March 2029 (or later)	6G specifications

16 supports all the use cases and requirements for 5G and conforms with the requirements of IMT-2020 [35]. Release 17 added enhancements such as the use of satellites and other non-terrestrial networks, and took the opportunity to remove some outdated terms from the specifications [36, 37]. 5G systems from Release 18 onwards are known as *5G-Advanced* [38], while work on a *sixth generation* (6G) is due to start in Release 20.

The specifications are also organized into several *series*, each of which covers a particular part of the system. Table 1.2 summarizes the contents of series 21–38, which contain all the specifications for UMTS, LTE and 5G, as well as specifications that are common with GSM. (Some other series numbers are used exclusively for GSM.) Although most series contain specifications for each of the 3GPP technologies, the 38 series is devoted to the 5G air interface and RAN, so it is an important source of information for this book.

When written in full, an example specification number is TS 23.501 v. 18.7.0. Here, TS stands for *Technical Specification*, which is the output from the 3GPP *work item* in which the specification was created. The series number is 23, 501 is the specification number within that series, 18 is the release number, 7 is the technical version number within that release and the final 0 is an editorial version number that is occasionally used for non-technical changes. 3GPP also produces *Technical Reports* (TR), which are the output from preliminary 3GPP *study items*. They are purely informative and have three-digit specification numbers beginning with 9, 8 or occasionally 7.

In a final division, each specification belongs to one of three *stages*, following a scheme that is defined by the ITU [39]. Stage 1 specifications define the service from the user's point of view and

Table 1.2 3GPP specification series used by UMTS, LTE and 5G.

Series	Scope
21	High-level requirements
22	Stage 1 service specifications
23	Stage 2 service and architecture specifications
24	Non-access stratum protocols
25	UMTS air interface and radio access network
26	Codecs
27	Data terminal equipment
28	Network management, including for 5G
29	Core network protocols
30	Programme management
31	UICC and USIM
32	Network management and charging
33	Security
34	UE test specifications
35	Security algorithms
36	LTE air interface and radio access network
37	Multiple radio access technologies
38	5G air interface and radio access network

lie exclusively in the 22 series. Stage 2 specifications define the system's high-level architecture and operation, and lie mainly (but not exclusively) in the 23 series. Finally, stage 3 specifications define all the functional details.

The stage 1 and 2 specifications are especially useful for acquiring a high-level understanding of the system. The most useful ones for 5G are TS 22.261, TS 23.501, TS 23.502 and TS 38.300 [40–43], which respectively cover the requirements, the system architecture, the high-level signalling procedures and the 5G air interface. Others include TS 23.503, TS 33.501, TS 37.340 and TS 38.401 [44–47], which cover policy control, security, dual connectivity with LTE and the 5G radio access network. There is, however, an important note of caution: those specifications are superseded later on and cannot be relied on for complete accuracy. Instead, the details should be checked in the relevant stage 3 specifications.

The individual specifications can be downloaded from 3GPP's specification numbering page or from its FTP server [48, 49]. The 3GPP website has summaries of the features in early releases [50], and later releases are summarized as technical reports [51–55].

1.4.2 Technical Performance Capabilities

Based on its earlier studies, the ITU prepared a set of performance requirements for IMT-2020, which it published in November 2017 [56]. In parallel, 3GPP published its own requirements for the 5G air interface in October 2016 [57, 58]. The 3GPP system is designed to be compliant with IMT-2020, so the two sets of requirements are very similar; they are summarized in Table 1.3.

Table 1.3 Technical performance requirements for 5G.

Parameter	LTE-Advanced	5G	Main 5G use case
Peak data rate	500 Mbps (UL)	10 Gbps (UL)	eMBB
	1000 Mbps (DL)	20 Gbps (DL)	
Typical spectral efficiency[a]	0.7–2.25 bits s^{-1} Hz^{-1} (UL)	1.6–6.75 bits s^{-1} Hz^{-1} (UL)	eMBB
	1.1–3 bits s^{-1} Hz^{-1} (DL)	3.3–9 bits s^{-1} Hz^{-1} (DL)	
Energy efficiency	n/a	Maximize	eMBB
Maximum UE speed	350 km h^{-1}	500 km h^{-1}	eMBB
User plane latency	5 ms	0.5 ms	URLLC
User plane reliability[b]	n/a	99.999%	URLLC
Maximum coupling loss[c]	n/a	143 dB	eMBB
	164 dB	164 dB	mMTC
Battery life[d]	10 yr	10–15 yr	mMTC
Connection density	60 000 km^{-2}	1 000 000 km^{-2}	mMTC

[a] Targets in various ITU performance environments.
[b] Delivery of a 32 byte packet within 1 ms.
[c] Data rate of 1 Mbps DL/30 kbps UL for eMBB and 160 bps for mMTC.
[d] Delivery of 200 uplink bytes per day, with a coupling loss of 164 dB and a 5 Wh battery.

The table also states the earlier requirements for LTE-Advanced, incorporating those for IMT-Advanced and, where relevant, NB-IoT [5, 59, 60].

Looking first at the peak data rate, the design targets for LTE-Advanced were 1000 Mbps on the downlink and 500 Mbps on the uplink. The design targets for 5G were far greater: 20 Gbps on the downlink and 10 Gbps on the uplink. The increase is mainly achieved by increasing the maximum bandwidth of each individual cell, from 20 MHz in the cases of LTE and LTE-Advanced, to an initial value of 400 MHz in the case of 5G. (We will discuss the peak data rates that the two systems can actually handle as part of Chapter 11.)

It cannot be stressed too strongly, however, that these peak data rates can only be reached in ideal conditions, for example if the bandwidth is very large and the received signal is very strong. A better measure is the typical *spectral efficiency*, which expresses the typical capacity of one cell per unit bandwidth, in units of bits s^{-1} Hz^{-1}. 5G was designed to support a spectral efficiency about three times greater than the original target for LTE-Advanced. (Once again, we will discuss the figures that are actually achieved as part of Chapter 11.)

In LTE networks, around 15% of the operating expenditure is for electrical power alone, with around 80% of that consumed by the base stations [61]. There is a risk of that situation worsening as the data rates increase, so there was a qualitative requirement to maximize the 5G network's *energy efficiency*, which expresses the capacity of one cell per unit of electrical power in units of bits s^{-1} W^{-1}.

The air interface's user plane latency is the time required for a data packet to travel between the mobile and the backhaul, provided that the air interface is not congested. (Additional delays can arise within the fixed network itself or in cases of congestion.) To handle URLLC applications, 5G supports an air interface latency as low as 0.5 ms and also requires the air interface to deliver packets within that time with a reliability as high as 99.999%.

The remaining requirements are applicable to machine-type communications and address the need to support a large number of machine-type devices with a long battery life. (The coupling loss is related to the range of the communication system and will be defined in Chapter 4.)

5G does not have to meet all of these requirements at the same time. Typically, an individual network slice will support some requirements but not others, for example the requirements for low latency and high reliability might be supported by a slice intended for URLLC, while those for enhanced coverage and long battery life might be supported by a slice for mMTC.

1.4.3 High-level Architecture

Architecturally, 5G builds on LTE by adding a number of new components, which are shown in Figure 1.6. The *5G core network* (5GC) contains new network functions that support NFV, SDN and network slicing. The *next-generation radio access network* (NG-RAN) introduces new interfaces and signalling procedures for the backhaul and for communications between nearby base stations, and supports RAN centralization and virtualization, edge computing and network slicing. Together, the two networks form the *5G system* (5GS). The 5G system does not interact with GSM or UMTS, with one exception that we will highlight in Section 1.5.

The 5G base station is known as a *next-generation Node B* (gNB). The gNB communicates with the mobile over an air interface known as the *New Radio* (NR), which supports millimetre-wave and multiple antenna communications.

Generically, an individual eNB or gNB is known as a *node*. As shown in Figure 1.6, an individual node can lie in either the E-UTRAN, the NG-RAN or both. Specifically, a node lies in the E-UTRAN if it supports the legacy LTE backhaul and is connected to the evolved packet core, and in the NG-RAN if it supports the new 5G backhaul and is connected to the 5G core.

1.4.4 Architectural Options

As implied by Figure 1.6, one of the important features of 5G is its close inter-operation with LTE. To help enumerate the different possibilities, the 5G specifications support a number of different architectural options, which were identified and numbered as part of the early 3GPP studies into 5G [62]. We will discuss these options more fully as part of Chapter 3, but

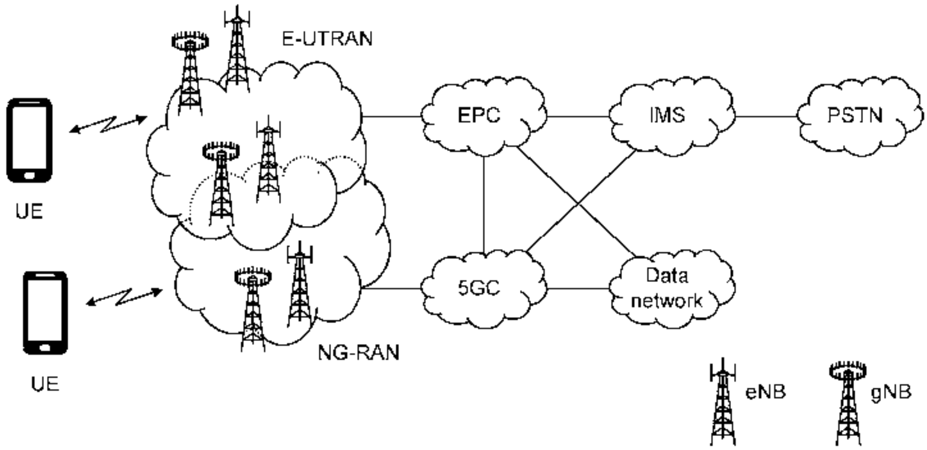

Figure 1.6 High-level architectures of LTE and 5G.

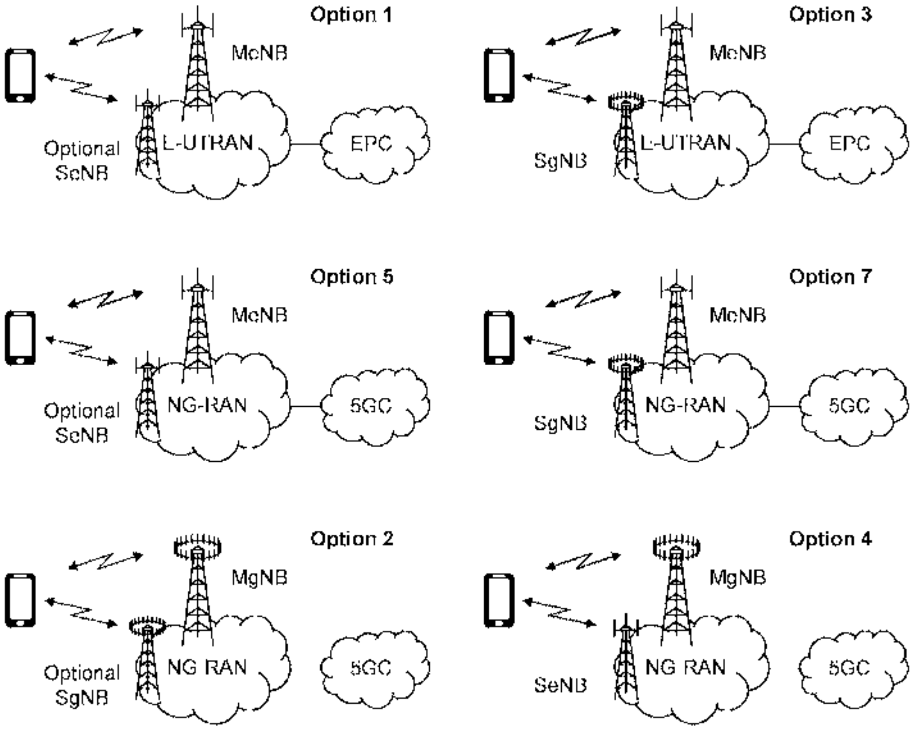

Figure 1.7 Architectural options for 5G.

they are summarized in Figure 1.7. An individual network operator can support one, some or all of these options at the same time.

Option 1 is a legacy LTE system that supports dual connectivity. Option 3 replaces the secondary eNB with a secondary gNB (SgNB), and is the one most commonly used by early roll-outs of 5G. The next two options introduce the 5G core network and the NG-RAN, using a master eNB and an optional secondary eNB in option 5, and a secondary gNB in option 7. In the last two options, the gNB takes on the role of a master gNB (MgNB), operating either alone or with a secondary gNB in option 2 and with a secondary eNB in option 4.

With the exception of option 1, each of these options uses aspects of 5G. However, some care is needed to distinguish the use of 5G in the air interface, the fixed network or both. On the air interface, the term *standalone* (SA) refers to architectures that use a single radio access technology, while *non-standalone* (NSA) refers to a mix of technologies [63]. With that definition, options 1 and 5 are standalone implementations of the LTE air interface, option 2 is a standalone implementation of the 5G New Radio, and options 3, 4 and 7 are all non-standalone. In the fixed network, option 3 is based on the LTE evolved packet system, while options 2, 4, 5 and 7 are based on the 5G system. Option 2 is the only one that uses 5G alone.

Because of that issue, parts of the book are only relevant to certain architectural options, and not to others. In an attempt to resolve any confusion, Table 1.4 summarizes how the material in this book maps onto the options listed above, with ticks denoting topics that are relevant and brackets denoting those that are relevant with restrictions or modifications. As an example, the architectural material in Chapters 2 and 3 mainly addresses the 5G system, but it also covers the core and radio access networks of LTE to meet the needs of architectural option 3.

Table 1.4 Mapping between the 5G architectural options and the contents of the book.

Chapters	Topics	Option 3	Option 5	Option 7	Option 2	Option 4
1	Introduction	✓	✓	✓	✓	✓
2-3	System architecture	(✓)	✓	✓	✓	✓
4-6	Wireless communications	✓	(✓)	✓	✓	✓
7-12	5G New Radio	✓		✓	✓	✓
13-19	5G system operation		✓	✓	✓	✓
20-23	5G-Advanced		✓	✓	✓	✓

1.5 Enhanced Mobile Broadband

1.5.1 eMBB Applications

We now return to the three use case families that we introduced in Section 1.2. The most fundamental is enhanced mobile broadband (eMBB), which addresses similar markets to LTE but with improved capabilities [13]. In particular, the eMBB use case calls for a higher data rate than LTE can provide in terms of the peak data rate in ideal conditions, the expected data rate in more typical conditions and the minimum in conditions of poor network coverage.

Applications are drawn from the traditional business and consumer markets. Examples include data downloads and real-time video in an indoor office environment, traffic hotspots such as shopping centres and other dense urban environments, public events such as football matches and concerts, and the provision of a consistent mobile broadband experience for consumers in rural areas and on public transport.

Some eMBB applications are controlled by the IP multimedia subsystem. Two in particular stand out. Firstly, the 5G system handles voice calls using VoIP alone, with the calls controlled by either a third-party application or the IP multimedia subsystem [64–66]. Release 16 introduces limited support for single radio voice call continuity, such that the network can convert an ongoing IMS voice call to a circuit-switched call, and can hand the mobile over to a 3G cell [67]. The mobile's communications with any external packet data networks are torn down, while there is no support for SRVCC to a 2G cell and no support for circuit-switched fallback at all. The second application involves public safety networks for the emergency services, which the 5G system supports from Release 17 [68–71].

1.5.2 The Market for eMBB

To illustrate the market for eMBB, Figure 1.8 shows the total traffic being handled by mobile telecommunication networks throughout the world in units of exabytes (10^{18} bytes, i.e. one billion gigabytes) per month. The data are from Ericsson [72, 73], and are based on measurements up to the third quarter of 2024 and forecasts thereafter. The networks are dominated by mobile broadband applications, with video making by far the biggest contribution. Noting the log scale on the vertical axis, the traffic levels have increased by a factor of over 100 since 2012, but the rate of growth has slowed considerably and is currently around 15–20% per year.

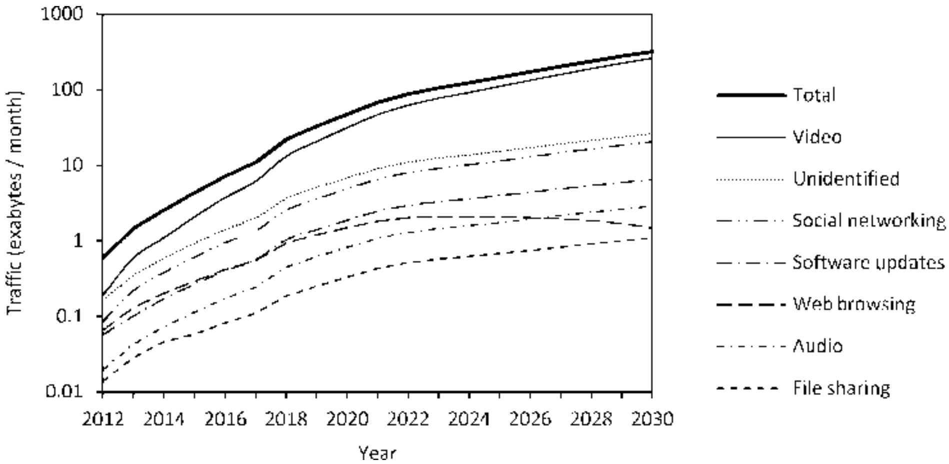

Figure 1.8 Monthly traffic due to different mobile telecommunication applications, with measurements up to the third quarter of 2024 and forecasts thereafter. *Source:* Data from Ericsson 2024.

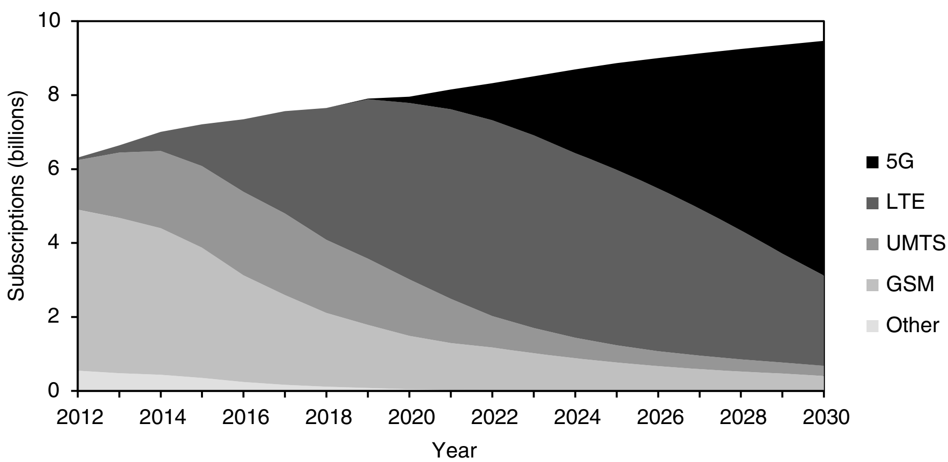

Figure 1.9 Numbers of subscriptions to different mobile telecommunication technologies, with measurements up to the third quarter of 2024 and forecasts thereafter. *Source:* Data from Ericsson 2024.

As the traffic levels have grown, there has been a shift towards devices that support the 5G New Radio. Nevertheless, developed markets have become saturated with smartphones and sales have stalled. To illustrate that, Figure 1.9 shows Ericsson's measurements and forecasts of the numbers of mobile subscriptions worldwide, with subscriptions for machine-type communications excluded. The total number of subscriptions overtook the world's population in 2018 because some subscriptions are inactive and because some users own more than one device. The number of unique subscribers in the third quarter of 2024 was about 6.6 billion.

To make matters worse, increasing levels of competition have reduced network operators' revenue per bit, while subscribers have also been making greater use of free WiFi access and of third-party services for voice and messaging. Together, these issues have caused network

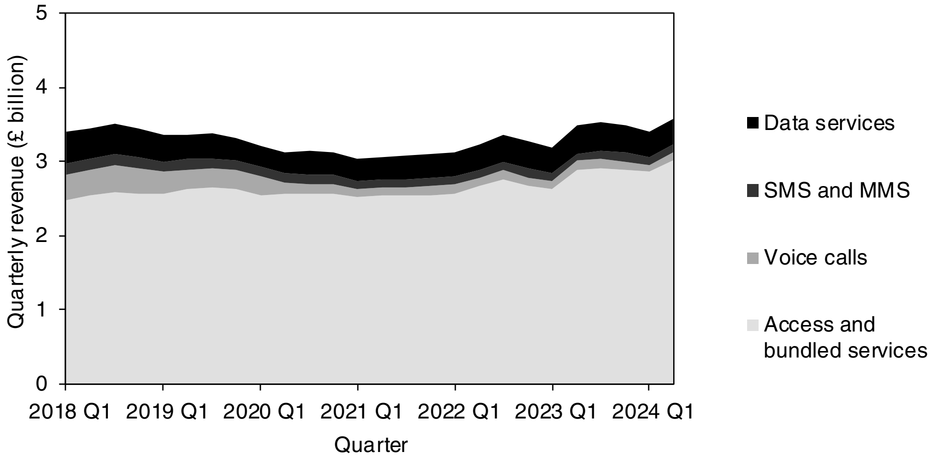

Figure 1.10 Quarterly revenue earned by UK mobile network operators up to the second quarter of 2024. *Source:* Data from Ofcom 2024.

operators' income to stall. To illustrate that point, Figure 1.10 shows estimates of the quarterly revenue earned by network operators in the UK up to the second quarter of 2024, which have been published by the UK regulator Ofcom [74]. (Ofcom's data for earlier years use a different accounting standard, so they are not directly comparable.)

Taking all of these issues into account, network operators have tended to conclude that the traditional consumer market is not a promising source of future revenue. Their response has been to look elsewhere.

1.6 Massive Machine-type Communications

1.6.1 mMTC Applications

The second use case is machine-type communications (MTC), also known as the *internet of things* (IoT) [14]. These terms both refer to wireless communications between autonomous machine-type devices without any direct human interaction. However, the first term is sometimes restricted to point-to-point communications, whereas the second implies that the devices communicate over the internet. In turn, the terms massive machine-type communications (mMTC) and *massive internet of things* (MIoT) both imply the aggregation and analysis of data from large numbers of connected devices.

The applications for MTC are wide-ranging. One example is wireless tracking devices, for example to report the physical location of goods in a warehouse, vehicles in a delivery system or cats in a local neighbourhood. Another is wireless sensors, such as motion-triggered security cameras, environmental sensors and strain gauges for bridges, buildings and other civil engineering structures. A third application is electronic health, including the reporting of medical information such as body temperature, blood pressure and glucose levels, to help permit independent living for the elderly and for those with long-term medical conditions.

MTC has different performance requirements from mobile broadband. The devices must be cheap and they require a low power consumption to ensure a long battery life. The devices must also communicate successfully if the received signal is very weak to support devices such as smart meters that might be installed deep inside a building. The data rate of an individual machine-type device is usually low, but the collective data rate can still be high. Furthermore, the associated signalling messages can overwhelm the network if they are not carefully controlled, for example if the devices try to re-establish communications following a network outage.

5G supports massive machine-type communications in two ways, which we will discuss as part of Chapter 21. Firstly, Release 16 integrates the eMTC and NB-IoT air interfaces into the 5G system, such that a machine-type device can communicate with the 5G core network through a base station that supports one of those technologies. Secondly, Release 17 introduces 5G *reduced capability* (RedCap) devices, which communicate using the 5G New Radio, but are simpler and cheaper than a normal 5G device. They are aimed at applications such as wireless video cameras, whose requirements exceed those of eMTC or NB-IoT but fall short of the underlying capabilities of 5G.

1.6.2 The Market for MTC

In 2024, there were around 20 billion machine-type devices worldwide. The majority used unlicensed technologies with a limited range of around 100 m, such as WiFi, Bluetooth and Zigbee. There were also around 4 billion devices using wide-area technologies, which are collectively known as *low power wide area* (LPWA) wireless. Two of those technologies operate in unlicensed spectrum, namely *Long Range Wide Area Network* (LoRaWAN) and *Sigfox*, but the majority make use of cellular communications in licensed spectrum. These are illustrated in Figure 1.11, which contains measurements and forecasts by Ericsson of the numbers of cellular IoT connections worldwide [72, 73].

The figure shows three main technology categories. Firstly, GPRS offers a basic set of capabilities that have long been used for cellular IoT, and which were subsequently enhanced by a technology known as *extended coverage GSM* (EC-GSM). There is also some limited use of 3G. The

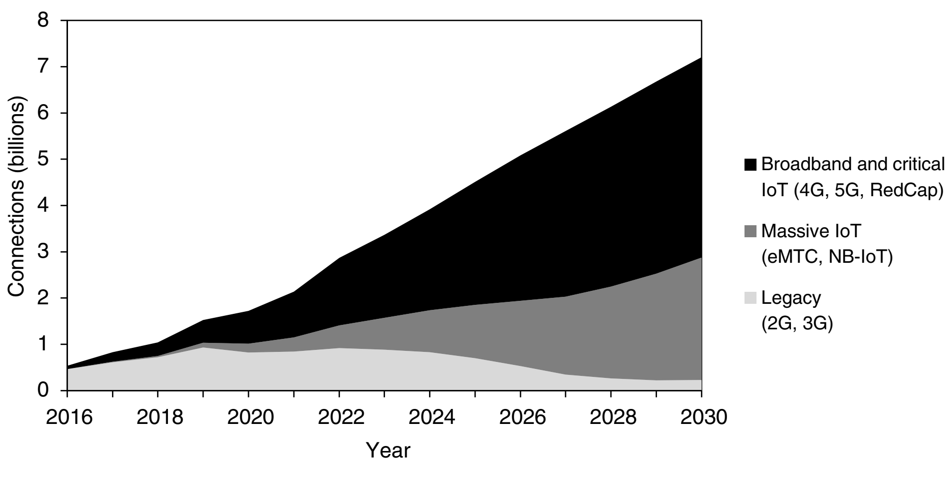

Figure 1.11 Numbers of cellular IoT devices using different mobile telecommunication technologies, with measurements up to the third quarter of 2024 and forecasts thereafter. *Source:* Data from Ericsson 2024.

second category encompasses eMTC and NB-IoT. The third covers a mix of 4G and 5G technologies with improved capabilities, including 5G RedCap devices, as well as the devices for low latency communication that we will discuss in Section 1.7.

1.7 Ultra-reliable Low-latency Communication

1.7.1 URLLC Applications

Ultra-reliable low-latency communication, also known as *critical communication* (CriC), is a family of use cases that are characterized by the need for very low latency, often in conjunction with very high reliability [15].

Latency is the time required to deliver packets between the mobile and an external server. LTE supports latencies of a few tens of milliseconds for non-roaming mobiles, which are low enough for conversational VoIP and video over IP. However, they are still too high for other applications, for example wireless industrial control systems, vehicle communications and extended reality. Meanwhile, the LTE air interface can deliver traffic with very high reliability, but only if the latency is high enough (tens or hundreds of milliseconds) for packets to be re-transmitted if they arrive incorrectly. It is far more difficult to deliver packets reliably with a lower latency, and a new air interface is required.

URLLC encompasses applications in which 5G can offer the greatest benefits over LTE, both because of their technical requirements and because of the opportunity they offer network operators to go beyond their traditional consumer markets. 5G supports URLLC through a number of enhancements to the air interface and the network architecture [75–77], which will be an ongoing theme throughout this book.

1.7.2 The Industrial Internet of Things

Low latency and high reliability come together in the concept of wireless industrial automation. This is often known as the *Industrial Internet of Things* (IIoT), and is sometimes depicted under the buzzword *Industry 4.0* as a fourth industrial revolution, which it is hoped will succeed the original industrial revolution of the late eighteenth and early nineteenth centuries, the introduction of mass-production around the start of the twentieth century and the introduction of computers from the 1950s. For example, a factory or larger industrial site might contain robots and other industrial machinery that are equipped with large numbers of sensors and actuators, and are governed in a closed-loop control system by a central server. In such a factory, wireless control brings several advantages over a network of fixed cables because it is cheaper and easier to install the system, to add new machinery and especially to reconfigure the factory in response to changes in customer demand. Although such a wireless control system might operate using unlicensed spectrum, a licensed system such as 5G can operate over longer ranges and with lower levels of interference, and hence with higher reliability.

Wireless industrial control systems have requirements that go beyond the basic capabilities of URLLC, and which also arise in applications such as smart energy distribution and live television production [78–84]. The most important is the need to synchronize all the devices to the same external clock, while others include the need to operate in private 5G networks that are not accessible to the general public and to integrate with older solutions that are based on local area networks. The 5G system supports these requirements from Release 16 and we will discuss them as part of Chapter 21.

1.7.3 Vehicle-to-everything Communication

Vehicle-to-everything (V2X) communication refers to the exchange of information between a road vehicle and the mobile telecommunication network, and with other vehicles and pedestrians that are nearby. It is best considered as an aspect of URLLC that has extra requirements of its own, in particular the need to handle fast-moving mobiles and to support device-to-device communications on the sidelink.

Vehicle communications can be carried out using unlicensed spectrum, for example using a variant of WiFi known as *dedicated short-range communication* (DSRC), which is defined in the standard IEEE 802.11p. As in the case of industrial IoT, however, cellular networks offer several advantages, in particular a higher transmit power and a lower level of interference, which lead to longer range and higher reliability. LTE was enhanced to offer basic support for vehicle communications as part of LTE-Advanced Pro, but that was limited by LTE's capabilities for data rate, latency and reliability. Such support is greatly improved by 5G and is implemented from Release 16 [85, 86]. We will address vehicle communications as part of Chapter 22.

1.7.4 Extended Reality and Media

Extended reality (XR) is an umbrella term for three types of service: *augmented reality* (AR) overlays additional information onto a natural scene; *virtual reality* (VR) generates an artificial scene by a process known as *rendering*; and *mixed reality* (MR) inserts artificial elements into a natural scene to give the illusion that they are the same. As a simple example, virtual reality headsets find consumer applications in immersive entertainment and online games, and require very low latencies to avoid feelings of nausea on the part of the user.

Extended reality services can trigger the introduction new types of media that go beyond the traditional audio and video. The most demanding are so-called *haptic data*, which are responsive over the timescales of human touch and require feedback within a few milliseconds to give a realistic impression to the user. Other examples might include text information for speech synthesis or augmented reality applications, and information about the surrounding environment such as its brightness, temperature or humidity.

The requirement for low latency triggers a need for coordination and synchronization among the different types of data stream. To address that issue, Release 18 introduces support for extended reality services that involve multiple streams of traffic, but which are limited to a single source and destination [87]. Release 19 extends that support to so-called *metaverse services*, which require coordination among multiple sources and multiple destinations [88, 89]. Even so, some extended reality applications require a combination of high data rate, low latency and high reliability that lies beyond the capabilities of 5G. Those applications will require a sixth-generation system, which we will address in Chapter 23.

References

1 Shannon, C.E. (1948). A mathematical theory of communication. *The Bell System Technical Journal* 27: 379–428. and 623–656.

2 Hampton, J.R. (2014). *Introduction to MIMO Communications*. Cambridge University Press.

3 ITU-R Recommendation M.1034-1 (1997) Requirements for the radio interface(s) for International Mobile Telecommunications 2000 (IMT-2000), February 1997.
4 3GPP (2024) 3GPP – The mobile broadband standard. www.3gpp.org (accessed 22 October 2024).
5 ITU-R Report M.2134-0 (2008) Requirements related to technical performance for IMT-Advanced radio interface(s), November 2008.
6 ITU-R Report M.2290-0 (2013) Future spectrum requirements estimate for terrestrial IMT, December 2013.
7 ITU-R Report M.2320-0 (2014) Future technology trends of terrestrial IMT systems, November 2014.
8 ITU-R Report M.2370-0 (2015) IMT traffic estimates for the years 2020 to 2030, July 2015.
9 ITU-R Report M.2376-0 (2015) Technical feasibility of IMT in bands above 6 GHz, July 2015.
10 ITU-R Recommendation M.2083-0 (2015) IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond, September 2015.
11 5G Americas (2016) Global organizations forge new frontier of 5G: 5G Americas 5G global update.
12 3GPP TR 22.891 (2016) Feasibility study on new services and markets technology enablers; Stage 1 (Release 14), September 2016.
13 3GPP TR 22.863 (2016) Feasibility study on new services and markets technology enablers – Enhanced mobile broadband; Stage 1 (Release 14), September 2016.
14 3GPP TR 22.861 (2016) Feasibility study on new services and markets technology enablers for massive internet of things; Stage 1 (Release 14), September 2016.
15 3GPP TR 22.862 (2016) Feasibility study on new services and markets technology enablers for critical communications; Stage 1 (Release 14), September 2016.
16 3GPP TR 22.864 (2016) Feasibility study on new services and markets technology enablers – Network operation; Stage 1 (Release 15), September 2016.
17 Chayapathi, R., Hassan, S.F., and Shah, P. (2017). *Network Functions Virtualization (NFV) with a Touch of SDN*. Addison Wesley.
18 ETSI (2012) Network functions virtualisation: An introduction, benefits, enablers, challenges & call for action. https://portal.etsi.org/NFV/NFV_White_Paper.pdf (accessed 22 October 2024).
19 ETSI GS NFV 002 (2014) Network functions virtualisation (NFV): Architectural framework, December 2014.
20 Nguyen, V.-G., Brunstrom, A., Grinnemo, K.-J. et al. (2017). SDN/NFV-based mobile packet core network architectures: a survey. *IEEE Communications Surveys and Tutorials* 19 (3): 1567–1602.
21 3GPP TS 28.500 (2024) Management concept, architecture and requirements for mobile networks that include virtualized network functions (Release 18), April 2024.
22 Goransson, P., Black, C., and Culver, T. (2016). *Software Defined Networks: A Comprehensive Approach*, 2e. Morgan Kaufmann.
23 Kalixylos, A. (2018). A survey and an analysis of network slicing in 5G networks. *IEEE Communications Standards Magazine* 2 (1): 60–65.
24 Afolabi, I., Taleb, T., Samdanis, K. et al. (2018). Network slicing and softwarization: a survey on principles, enabling technologies, and solutions. *IEEE Communications Surveys and Tutorials* 20 (3): 2429–2453.
25 O-RAN Alliance (2024), O-RAN Alliance e.V https://www.o-ran.org (accessed 22 October 2024).
26 ETSI (2014) Mobile-edge computing – Introductory technical white paper. https://portal.etsi.org/Portals/0/TBpages/MEC/Docs/Mobile-edge_Computing_-_Introductory_Technical_White_Paper_V1%2018-09-14.pdf (accessed 22 October 2024).
27 ETSI GS MEC 003 (2024) Multi-access edge computing (MEC); Framework and reference architecture, April 2024.

28 ETSI (2018) MEC in 5G networks. https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf (accessed 22 October 2024).
29 Taleb, T., Samdanis, K., Mada, B. et al. (2017). On multi-access edge computing: a survey of the emerging 5G network edge cloud architecture and orchestration. *IEEE Communications Surveys and Tutorials* 19 (3): 1657–1681.
30 5G Americas (2019) Advanced antenna systems for 5G.
31 Britannica (2024) Artificial Intelligence. https://www.britannica.com/technology/artificial-intelligence (accessed 22 October 2024).
32 3GPP TS 28.105 (2024) Management and orchestration; Artificial Intelligence/Machine Learning (AI/ML) management (Release 18), September 2024.
33 3GPP (2024) The 3GPP's system of parallel releases. https://www.3gpp.org/specifications-technologies/releases (accessed 22 October 2024).
34 3GPP (2024) Release 15. https://www.3gpp.org/specifications-technologies/releases/release-15 (accessed 22 October 2024).
35 ITU (2020) ITU completes evaluation for global affirmation of IMT-2020 technologies. https://www.itu.int/en/mediacentre/Pages/pr26-2020-evaluation-global-affirmation-imt-2020-5g.aspx (accessed 22 October 2024).
36 3GPP (2021) Initiative to remove non-inclusive terms. https://www.3gpp.org/news-events/3gpp-news/inclusivity (accessed 22 October 2024).
37 3GPP TR 21.801 (2024) Specification drafting rules (Release 18), March 2024, Sections 4.8, K.
38 3GPP (2024) Release 18. https://www.3gpp.org/specifications-technologies/releases/release-18 (accessed 22 October 2024).
39 ITU-T Recommendation I.130 (1988) Method for the characterization of telecommunication services supported by an ISDN and network capabilities of an ISDN, November 1988.
40 3GPP TS 22.261 (2024) Service requirements for the 5G system; Stage 1 (Release 18), September 2024.
41 3GPP TS 23.501 (2024) System architecture for the 5G system (5GS); Stage 2 (Release 18), September 2024.
42 3GPP TS 23.502 (2024) Procedures for the 5G system (5GS); Stage 2 (Release 18), September 2024.
43 3GPP TS 38.300 (2024) NR; NR and NG-RAN overall description; Stage 2 (Release 18), September 2024.
44 3GPP TS 23.503 (2024) Policy and charging control framework for the 5G system (5GS); Stage 2 (Release 18), September 2024.
45 3GPP TS 33.501 (2024) Security architecture and procedures for 5G system (Release 18), September 2024.
46 3GPP TS 37.340 (2024) Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Multi-connectivity; Stage 2 (Release 18), September 2024.
47 3GPP TS 38.401 (2024) NG-RAN; Architecture description (Release 18), September 2024.
48 3GPP (2024) Version numbering scheme. https://www.3gpp.org/specifications-technologies/specifications-by-series/version-numbering-scheme (accessed 22 October 2024).
49 3GPP (2024) ftp://ftp.3gpp.org (accessed 22 October 2024).
50 3GPP (2024) ftp://ftp.3gpp.org/Information/WORK_PLAN/Description_Releases (accessed 22 October 2024).
51 3GPP TR 21.914 (2018) Release 14 description; Summary of Rel-14 work items (Release 14), May 2018.
52 3GPP TR 21.915 (2019) Release 15 description; Summary of Rel-15 work items (Release 15), September 2019.

53 3GPP TR 21.916 (2022) Release 16 description; Summary of Rel-16 work items (Release 16), June 2022.
54 3GPP TR 21.917 (2023) Release 17 description; Summary of Rel-17 work items (Release 17), January 2023.
55 3GPP TR 21.918 (2024) Release 18 description; Summary of Rel-18 work items (Release 18), September 2024.
56 ITU-R Report M.2410-0 (2017) Minimum requirements related to technical performance for IMT-2020 radio interface(s), November 2017.
57 3GPP TR 38.913 (2024) Study on scenarios and requirements for next generation access technologies (Release 18), March 2024.
58 3GPP TS 22.261 (2024) Service requirements for the 5G system; Stage 1 (Release 18), September 2024, Section 7.
59 3GPP TR 36.913 (2024) Requirements for further advancements for Evolved Universal Terrestrial Radio Access (E-UTRA) (LTE-Advanced) (Release 18), March 2024.
60 3GPP TR 45.820 (2015) Cellular system support for ultra-low complexity and low throughput internet of things (CIoT) (Release 13), November 2015.
61 Nokia (2016) 5G network energy efficiency: Massive capacity boost with flat energy consumption. https://gsacom.com/paper/5g-network-energy-efficiency-nokia-white-paper (accessed 22 October 2024).
62 3GPP TR 38.801 (2017) Study on new radio access technology: Radio access architecture and interfaces (Release 14), March 2017, Section 7.2.
63 GSMA (2018) Road to 5G: Introduction and migration.
64 3GPP TS 22.173 (2024) IP multimedia core network subsystem (IMS) multimedia telephony service and supplementary services; Stage 1 (Release 18), March 2024.
65 3GPP TR 22.823 (2018) Study on enhancements to IMS for new real time communication services; Stage 1 (Release 16), June 2018.
66 3GPP TR 22.873 (2021) Study on evolution of IMS multimedia telephony service (Release 18), June 2021.
67 3GPP TR 23.756 (2018) Study for single radio voice continuity from 5GS to 3G (Release 16), September 2018.
68 3GPP TS 22.179 (2023) Mission Critical Push To Talk (MCPTT); Stage 1 (Release 18), June 2023.
69 3GPP TS 22.280 (2024) Mission Critical Services Common Requirements (MCCoRe); Stage 1 (Release 18), December 2023.
70 3GPP TS 22.281 (2024) Mission critical video services (Release 18), March 2024.
71 3GPP TS 22.282 (2024) Mission critical data services (Release 18), March 2024.
72 Ericsson (2024) Ericsson mobility report, November 2024. https://www.ericsson.com/en/reports-and-papers/mobility-report (accessed 28 November 2024).
73 Ericsson (2024) Ericsson mobility visualizer. https://www.ericsson.com/en/reports-and-papers/mobility-report/mobility-visualizer (accessed 28 November 2024).
74 Ofcom (2024) Telecommunications market data update Q2 2024. www.ofcom.org.uk/phones-and-broadband/telecoms-infrastructure/telecommunications-market-data-update-q2-2024 (accessed 28 November 2024).
75 3GPP TS 23.501 (2024) System architecture for the 5G system (5GS); Stage 2 (Release 18), September 2024, Section 5.33.
76 3GPP TR 28.832 (2024) Study on management aspects of Ultra-Reliable and Low Latency Communications (URLLC) (Release 18), June 2023, Section 5.1.

77 3GPP TS 38.300 (2024) NR; NR and NG-RAN overall description; Stage 2 (Release 18), September 2024, Section 16.1.

78 3GPP TR 22.821 (2018) Feasibility study on LAN support in 5G (Release 16), June 2018.

79 3GPP TR 22.804 (2020) Study on communication for automation in vertical domains (Release 16), July 2020.

80 3GPP TR 22.832 (2021) Study on enhancements for cyber-physical control applications in vertical domains; Stage 1 (Release 17), March 2021.

81 3GPP TS 22.104 (2024) Service requirements for cyber-physical control applications in vertical domains; Stage 1 (Release 18), June 2024.

82 3GPP TR 22.867 (2021) Study on 5G smart energy and infrastructure (Release 18), December 2021.

83 3GPP TR 22.827 (2019) Study on audio-visual service production; Stage 1 (Release 17), December 2019.

84 3GPP TS 22.263 (2024) Service requirements for video, imaging and audio for professional applications (VIAPA); Stage 1 (Release 18), March 2024.

85 3GPP TR 22.886 (2018) Study on enhancement of 3GPP support for 5G V2X services (Release 16), December 2018.

86 3GPP TS 22.186 (2024) Enhancement of 3GPP support for V2X scenarios; Stage 1 (Release 18), March 2024.

87 3GPP TR 22.847 (2022) Study on supporting tactile and multi-modality communication services; Stage 1 (Release 18), March 2022.

88 3GPP TR 22.856 (2024) Feasibility study on localized mobile metaverse services (Release 19), December 2023.

89 3GPP TS 22.156 (2024) Mobile metaverse services; Stage 1 (Release 19), March 2024.