Radio fundamentals for cellular networks

White paper

ROHDE&SCHWARZ

Table of contents

Cellular technologies have advanced from first generation (1G) analog technologies to advanced high-performance fourth generation (4G) and fifth generation (5G) systems in just four decades. Despite the increase in complexity of wireless standards and devices, cellular technologies maintain a set of common principles that form the basis behind the design of cellular systems. In this white paper, we explore these basic principles and examine the underlying technologies that lay the foundation for today and future cellular systems.

Introduction

Cellular networks enable devices such as smartphones and internet of things (IoT) devices to communicate wirelessly. Cellular technologies have advanced from first generation (1G) analog technologies to advanced high-performance fourth generation (4G) and fifth generation (5G) systems in just about four decades. While 1G cellular technologies have disappeared, 2G technologies are rapidly being replaced by newer generations of technologies. 3G and 4G cellular technologies are widely deployed around the world. And 5G technologies have started to appear and will continue to be deployed through 2030.

Despite the increase in sophistication of wireless standards and devices over the years, cellular technologies maintain a set of common principles that form the basis behind the design of cellular systems. Certain principles are highly likely to be incorporated into 6G systems, whatever that standard turns out to be in the future. Certainly, the implementation of these underlying principles will vary from one standard to another and sometimes even within revisions of a given standard. We review these basic principles in this paper, and in some cases provide more explicit details by comparing 3G systems and 4G systems.

Section 1 introduces the composition of the network: the radio access network where the signals are processed and managed; the core network to facilitate establishment of an end-to-end logical link between the wireless device and an external entity such as a web server; and a services network that manages applications running on the cellular network. We discuss how the information of multiple users occupies the spectrum depending on the multiple access technique (sharing of the spectrum through time, frequency, space or coding separation) and the availability of spectrum to accommodate the base station to handset link (downlink) and the handset to the base station link (uplink). The different frequency bands that cellular systems can use are summarized along with the characteristics of those operating bands.

In the second section, we discuss the radio interface and describe the protocols implemented between the base station and the handset, the composition of a generic cellular transceiver and some of the emerging architecture trends in the implementation of the radio access network. Two examples of physical layer access are described: code division multiple access (CDMA), which is used in 3G systems, and orthogonal frequency division multiple access (OFDMA), which is used in 4G and 5G systems. Some of the features of 3G systems are discussed, for instance the ability to share a channel with multiple users, combining of multipath signals and soft handover and coding principles to overcome errors introduced by the channel. Likewise, some of the features of 4G and 5G systems are discussed, including the ability to share spectrum with multiple users, highly efficient coding schemes and the mechanism for dealing with multipath signals.

In section 3, we talk about processes that occur in any cellular network, such as how the handset can find a cell site (locate and extract the needed information for contacting the base station), inform a cell site that the handset is in the cell's coverage region and that it is a legitimate user, approach for the handset and the base station to exchange data, and processes for the handset to select and connect to another base station while it is moving in and out of the coverage areas of different cell sites (handover).

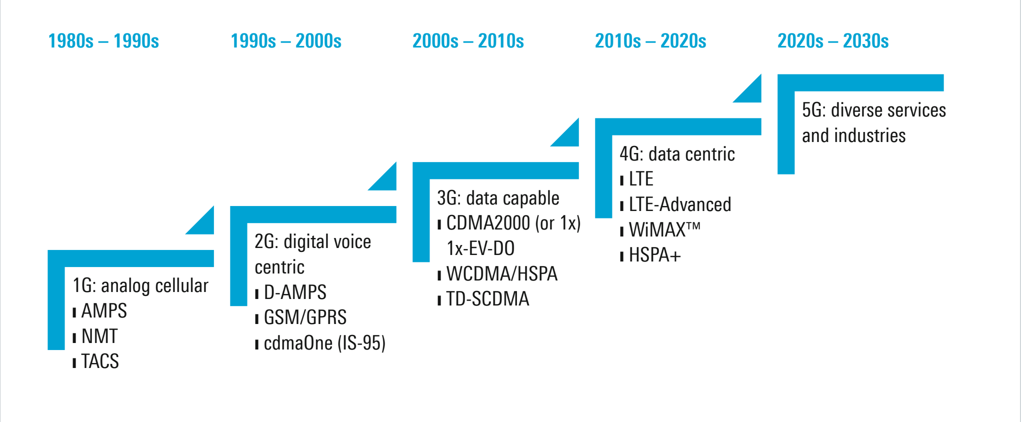
The last section summarizes key points and talks about emerging trends in cellular systems.

1 Making communications wireless

1.1 Evolution of cellular standards

Cellular networks enable devices such as smartphones and internet of things (IoT) devices to communicate wirelessly. Cellular technologies have advanced from first generation(1G) analog technology to advanced high-performance fourth generation (4G) and fifth generation (5G) systems in just about four decades [1]. While 1G cellular technologies have disappeared, 2G technologies are gradually being replaced by newer generations of technologies. 3G and 4G cellular technologies are widely deployed around the world. And 5G technologies have begun to appear in 2018.

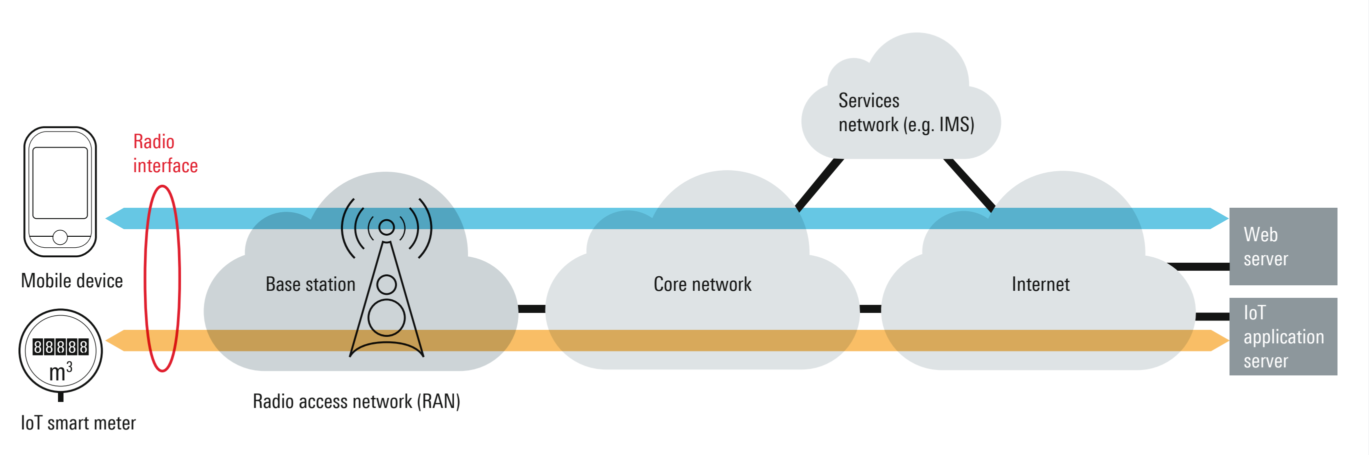
Evolution of cellular standards – main commercial deployments



Though the exact network architecture differs from one generation to another, a typical cellular network consists of a radio access network (RAN), a core network (CN) and a services network as shown in Fig. 1 [1]. The RAN contains base stations (BS) that communicate with the wireless devices using radio frequency (RF) signals, and it is this interface between the base station and the devices that is the primary subject of this paper. The RAN allocates radio resources to the devices to make wireless communications a reality. The CN performs functions such as user authentication, service authorization, security activation, IP address allocation and setup of suitable links to facilitate the transfer of user traffic such as voice and video. The services network includes operator-specific servers and IP multimedia subsystem (IMS) to offer a variety of services to the wireless subscriber, including voice calls, text messages (SMS) and video calls.

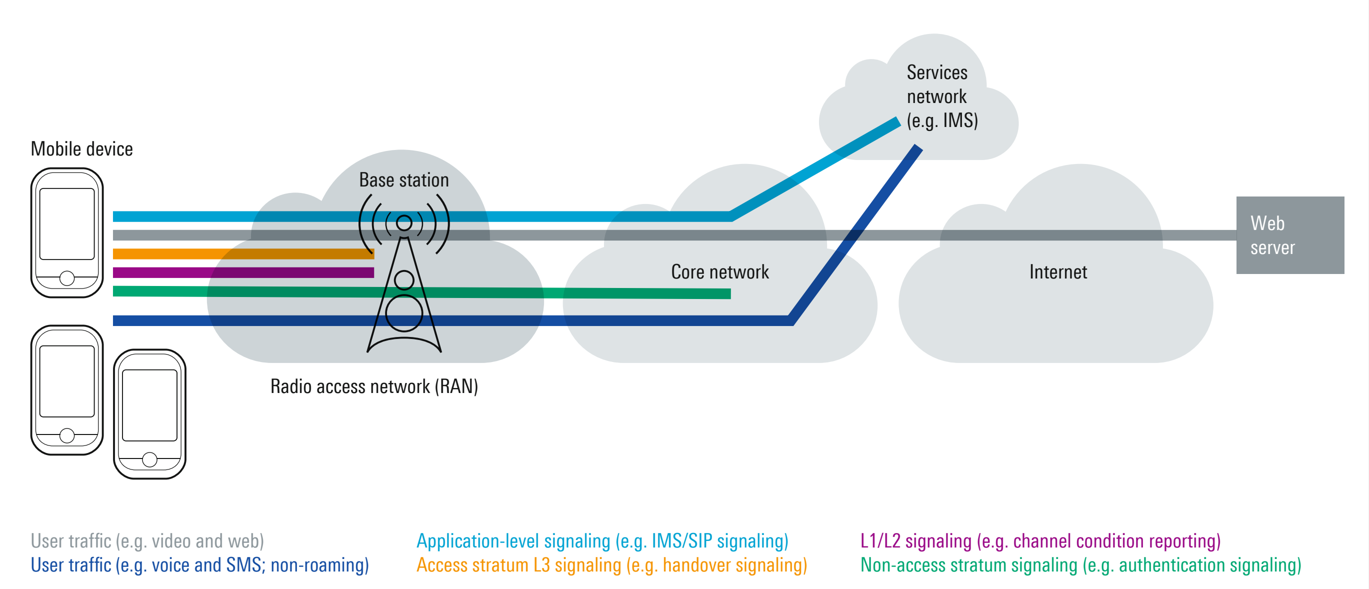
The cellular network interfaces with external networks such as the public switched telephone network (PSTN), the internet, enterprise networks and wireless fidelity (Wi-Fi) networks. The cellular network's connectivity with the internet allows wireless subscribers to access over-the-top (OTT) services such as YouTube videos, and the cellular network's connectivity with enterprise networks allows wireless subscribers to securely access private enterprise networks. Auxiliary systems (not shown in the diagram) such as operations support systems (OSS) and business support systems (BSS) help manage RAN, CN and services.

Fig. 1: Cellular network: a high-level view



Wireless devices exchange more than just the user traffic with the cellular network or wireless network [1]. In addition to user traffic such as voice, emails and videos, the devices and the cellular network exchange signaling messages. Signaling messages help set up voice calls and data sessions and carry out auxiliary functions such as authentication of user devices. Types of signaling include application-level signaling, cellular technology-specific signaling such as non-access stratum (NAS) and access stratum (AS) signaling, and lower layer signaling on the air interface. Protocols defined by standardization bodies such as 3GPP [2] and IETF [3] facilitate such information exchange between the device and the network. Fig. 2 shows the path traversed by different types of information in the cellular network as they are transferred between the endpoints of the communications link.

Fig. 2: Types of information passing through the cellular network


Services
network
(e.g. IMS)
Mobile device
Base station
Core network
Internet
Web
server
Radio access network (RAN)
User traffic (e.g. video and web)
User traffic (e.g. voice and SMS; non-roaming)
Application-level signaling (e.g. IMS/SIP signaling)
Access stratum L3 signaling (e.g. handover signaling)
L1/L2 signaling (e.g. channel condition reporting)
Non-access stratum signaling (e.g. authentication signaling)

[1] While this paper uses the term cellular network and wireless network interchangeably, the cellular network is one example of wireless networks. Examples of non-cellular wireless networks include Wi-Fi, Bluetooth® and satellite based networks.

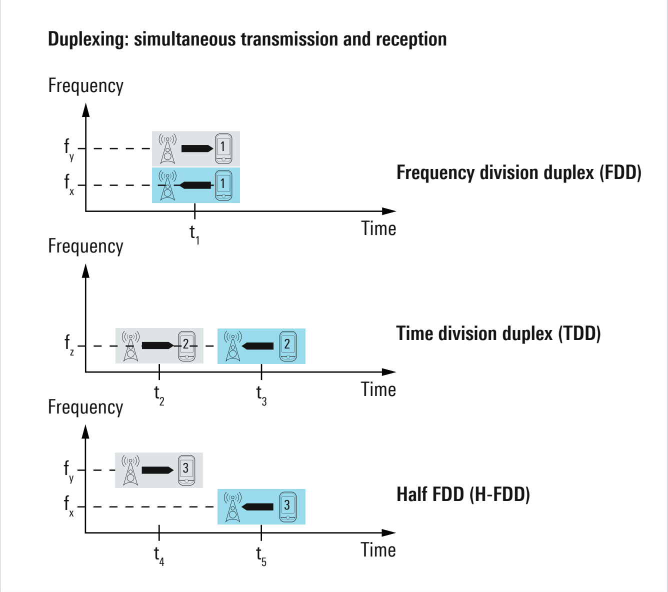
[2] 3GPP: 3rd Generation Partnership Project. 3GPP defined specifications for 3G technologies such as universal mobile telecommunication system (UMTS) and high-speed packet access (HSPA) and 4G standards such as long term evolution (LTE). 3GPP is now defining 5G standards.

[3] IETF, the Internet Engineering Task Force, has defined specifications for numerous protocols, including internet protocol (IP) and session initiation protocol (SIP).

When a user watches a video from a web server (e.g. YouTube), user traffic travels through the data network (e.g. the internet), the core network and the radio access network. For an IMS based voice call or SMS between two smartphones, the user traffic passes through the radio access network, the core network, the IMS network, the core network and the radio access network. Application-level signaling such as session initiation protocol (SIP) signaling between the device and the IMS network helps set up and tear down sessions such as voice calls. The device and the core network exchange technology-specific, non-access stratum (NAS) signaling messages. NAS signaling helps with functions such as authentication of the user by the network, authentication of the network by the device and activation of security. Layer 3 (L3) access stratum signaling between the device and the radio access network involves technology-specific signaling to support procedures such as radio interface configuration of the device for device to radio access network communications and handover of the device from one base station to another. The device and the radio access network exchange layer 1/layer 2 (L1/L2) signaling on the radio interface to facilitate reporting of radio channel conditions by the device to the radio access network and allocation of radio resources to the device by the radio access network. As shown in Fig. 2, a variety of information passes through the radio interface between the device and the radio access network.

Let us discuss communications between the device and the radio access network. A technique called duplexing allows the device or the base station to simultaneously transmit and receive information. Fig. 3 illustrates duplexing techniques.

Fig. 3: Duplexing in a cellular network



The communications link from the device to the base station is called the uplink or the reverse link, and the communications link from the base station to the device is called the downlink or the forward link. Duplexing allows the device and the base station to simultaneously send information on the one link while receiving information on the other link. Duplexing facilitates bidirectional and realtime transfer of information. Two basic duplexing methods are frequency division duplex (FDD) and time division duplex (TDD). A special case of FDD is half-duplex FDD (H-FDD).

In FDD, one part of the frequency spectrum is used for the uplink and a different part of the frequency spectrum for the downlink. From the device perspective, uplink transmission and downlink reception can occur at exactly the same time. From the base station perspective, downlink transmission and uplink reception can occur at exactly the same time. Paired spectrum with separate downlink spectrum and uplink spectrum is needed for FDD. The device and the base station also need more complex RF and DSP processing capability (e.g. simultaneously operating transmitter and receiver) for FDD.

In TDD, the same unpaired frequency spectrum is used for the uplink and the downlink. The uplink exists at one instant, and the downlink exists at a different instant. Since the switching between the uplink and the downlink is carried out rapidly (e.g. on the order of milliseconds before 5G or even tens of microseconds in 5G), the uplink and the downlink are considered "simultaneous" for all practical purposes.

TDD is simpler and less expensive than FDD from the device design perspective. However, interference is easier to manage with FDD due to the separation of the uplink and the downlink in the frequency domain. The uplink channel bandwidth and the downlink channel bandwidth tend to be identical in FDD due to the symmetric paired spectrum allocation by the governments, although cellular technologies often allow different aggregate channel bandwidths in the downlink and the uplink for FDD (e.g. 20 MHz in the downlink and 10 MHz in the uplink). Since TDD shares the same spectrum between the downlink and the uplink, the ratio of the downlink and the uplink can potentially be chosen to match overall traffic volume in the downlink and the uplink. For example, more time can be allocated to the downlink than the uplink when more data needs to be transferred in the downlink than in the uplink. FDD systems are very popular, but TDD systems are quickly gaining in popularity because of the easier availability of unpaired spectrum at higher frequencies.

Half-duplex FDD (H-FDD) can be viewed as a special case of FDD. Like FDD, H-FDD uses different chunks of spectrum for the uplink and downlink. However, at the device, only one link is active at an instant in time. Therefore, with H-FDD, the device either transmits in the uplink using the uplink spectrum or receives in the downlink using the downlink spectrum at a given instant. H-FDD is from the device perspective only; it is common for the base station to use regular FDD. The base station has both the downlink and the uplink active at the same time and can serve FDD devices and H-FDD devices. The base station ensures that the uplink and the downlink do not occur at the same time for a given H-FDD device. H-FDD simplifies device operation and reduces the cost of the user's device.

While duplexing allows simultaneous (or "almost" simultaneous) transmission and reception for a given entity such as the device and the base station, multiple access allows multiple devices to access and use the network at the same time through suitable sharing of radio resources. Fig. 4 depicts a simplified view of multiple access techniques commonly used in cellular networks[4].

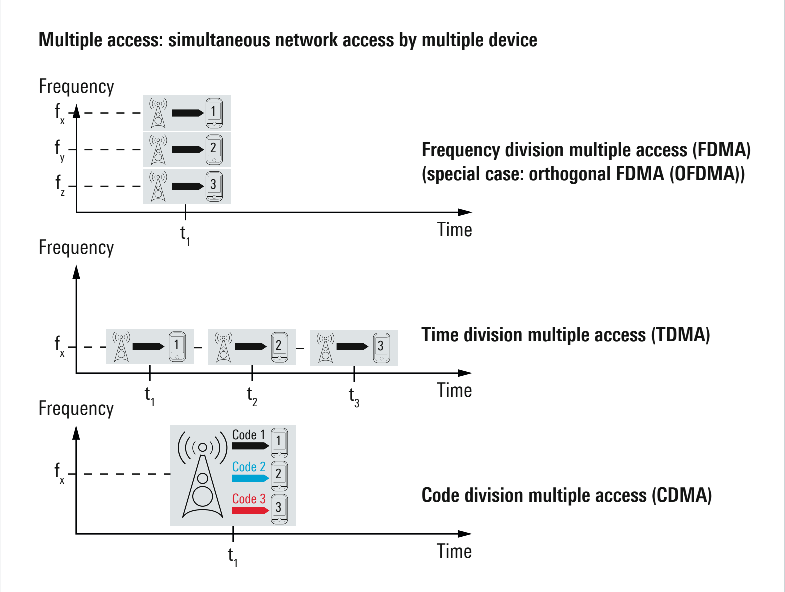
Frequency division multiple access (FDMA) allows multiple devices to access the network using different frequency channels. For example, three different users are assigned three distinct frequency channels (e.g. f_x, f_y and f_z in Fig. 4). An adequate guard band is designed between adjacent frequency channels to minimize interference between adjacent frequency channels. FDMA was widely used in 1G analog cellular networks. Modern 4G and 5G digital cellular networks use a sophisticated version of FDMA called orthogonal frequency division multiple access (OFDMA) where the frequency channels allocated to different devices are orthogonal to one another to achieve high spectral efficiency.

Time division multiple access (TDMA) allows multiple devices to access the network using different timeslots of a given frequency channel (e.g. frequency f_x in Fig. 4). TDMA has been widely used in 2G digital cellular networks (e.g. GSM, the global system for mobile communications systems).

[4] Thanks to advanced antenna techniques available in 4G and especially in 5G, space division multiple access (SDMA) is possible in addition to the traditional FDMA, TDMA and CDMA. SDMA is discussed in section 2.

Code division multiple access (CDMA) involves the use of a wideband frequency channel with different users using different orthogonal codes. Such orthogonal codes eliminate or minimize interference among users. For example, three different users can use three distinct orthogonal codes (e.g. codes 1, 2 and 3 in Fig. 4) to access the network simultaneously. 3G networks such as the universal mobile telecommunications system (UMTS) and 1x use CDMA.

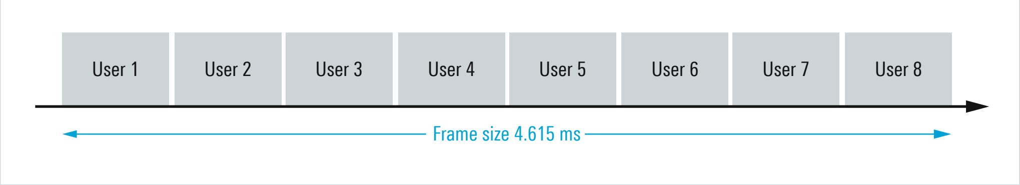
Fig. 4: Traditional multiple access in a cellular network


Multiple access: simultaneous network access by multiple device
Frequency
f_x
f_y
f_z
1
2
3
t_1
Time
Frequency division multiple access (FDMA) (special case: orthogonal FDMA (OFDMA))
Frequency
f_x
1
2
3
t_1
t_2
t_3
Time
Time division multiple access (TDMA)
Frequency
f_x
Code 1
Code 2
Code 3
1
2
3
t_1
Time
Code division multiple access (CDMA)

TDMA in practice: GSM

TDMA is one of the features pioneered in GSM. TDMA in GSM splits time up into eight timeslots for the uplink and downlink. Therefore, one 200 kHz channel of GSM can support eight users, each user in a 577 μs slot that comprises a 4.615 ms TDMA frame. An uplink frame is shown in Fig. 5. A similar allocation is made in the downlink frame.

Fig. 5: Frame of a GSM uplink signal



The data in each timeslot represents control information or speech and is encoded using convolution coding, puncturing (elimination of coded redundancy to reduce the number of data bits such that data fits in a frame), interleaving (to combat long channel fades that otherwise would create a continuous flow of channel induced errors) and differential encoding (to reduce ambiguity of mapping data to modulation symbols). Frames are constructed to provide the location of the physical channel (true information bearing signal) that carries a specific logical channel (abstracted label for information) or the structure that can provide information for signaling messages and to facilitate encryption.

Which is better TDMA or CDMA?

If one goes back to the basic concepts of information theory, it has been shown by Shannon-Hartley that the channel capacity C in bit/s has an upper boundary given by

$$C = B\log_2\left(1+\frac{S}{N}\right)$$

where B is the bandwidth in Hz, S is the averaged received signal power in that bandwidth in watts, and N is the average power in watts of the noise and interference within the bandwidth. Note that this expression does not depend on whether the signal is TDMA or CDMA. Relative differences in the achieved capacity are due to practical deployment issues such as ways to reduce interference between nearby signal sources or even in the implementation of the power amplifier.

Standardization bodies such as the 3rd Generation Partnership Project (3GPP) define the frequency spectrum used for communications between the device and the base station. Table 1 gives examples of the frequency bands for FDD and TDD. More FDD bands are defined for lower frequencies (e.g. below 6 GHz), while more TDD bands are defined for higher frequencies (e.g. millimeterwave spectrum). As cellular technologies evolve from one generation to the next, newer frequency bands are often defined to deploy newer generations of technologies since legacy (older generation) technologies often coexist with newer generation technologies for some years (e.g. one or two decades). A comprehensive list of frequency bands can be found in [2] and [3]. Newer bands (e.g. band 46 and higher) often tend to be TDD since unpaired TDD spectrum is easier to find than paired FDD spectrum.

Table 1: Example frequency bands for cellular communications

3GPP band	Type	Uplink frequency range	Downlink frequency range
n8 (GSM900)	FDD	880 MHz to 915 MHz	925 MHz to 960 MHz
2 (PCS 1900)	FDD	1850 MHz to 1910 MHz	1930 MHz to 1990 MHz
4 (AWS)	FDD	1710 MHz to 1755 MHz	2110 MHz to 2155 MHz
5 (850, cellular)	FDD	824 MHz to 849 MHz	869 MHz to 894 MHz
12 (lower 700)	FDD	699 MHz to 716 MHz	729 MHz to 746 MHz
13 (upper 700)	FDD	777 MHz to 787 MHz	746 MHz to 756 MHz
41 (TDD 2600)	TDD	2496 MHz to 2690 MHz	2496 MHz to 2690 MHz
46 (unlicensed)	TDD	5150 MHz to 5925 MHz	5150 MHz to 5925 MHz
n78 (5G NR, sub-6 GHz)	TDD	3300 MHz to 3800 MHz	3300 MHz to 3800 MHz
n257 (5G NR, mmW)	TDD	26 500 MHz to 29 500 MHz	26 500 MHz to 29 500 MHz
n260 (5G NR, mmW)	TDD	37 000 MHz to 40 000 MHz	37 000 MHz to 40 000 MHz

In Table 1, band 2 (PCS 1900 band) is an FDD band with an uplink spectrum of 1850 MHz to 1910 MHz and downlink spectrum of 1930 MHz to 1990 MHz. Initial 2G digital cellular technologies widely used band 2. 1G cellular technologies used band 5 (the 850 MHz cellular band). Initial 4G long term evolution (LTE) deployments in the U.S. were in the 700 MHz frequency bands 12 and 13. Band 41 has been popular for LTE TDD deployments. In a TDD band, the uplink frequency range is identical to the downlink frequency range because TDD makes use of the unpaired spectrum. While 3GPP primarily focuses on licensed spectrum, some unlicensed spectrum is also available for use in band 46. Band 46 is typically shared between cellular technologies and Wi-Fi. 5G New Radio (NR) technologies use sub-6 GHz and even up to mmW frequency bands such as FDD band n78 and TDD bands n257 and n260[5]. Many of the newer bands for cellular systems are shared bands with legacy spectrum users, and this trend is expected to continue and likely accelerate in the future.

[5] The letter “n” in the frequency band name represents New Radio, which implies a 5G frequency band.

1.2 GSM: A historical perspective

While this paper focuses on 3G, 4G and 5G cellular technologies, let us take a quick look at the most dominant second-generation technology: global system for mobile telecommunications (GSM). GSM was originally created to replace numerous first-generation analog cellular technologies with a pan-European system. However, GSM has since evolved to become the most dominant 2G technology, spanning all the continents and more than 160 countries around the globe, justifying the name Global.

Figure 2 illustrates the overall network architecture. In GSM, TDMA-based radio access network and a circuit-switched core network is used. The GSM RAN consists of base station controllers, with each BSC controlling a few hundred base stations. The mobile stations and the BS communicate using the TDMA-based air interface. TDMA is described in conjunction with Fig. 4, followed by a brief discussion of the GSM air interface in section 1.1.

The popular subscriber identity module (SIM) concept was also introduced by GSM. The SIM card stores the subscription profile of the user and personal information such as a list of contacts. The user can make use of the same SIM card with any suitable physical mobile device.

In 2.5G general packet radio service (GPRS, an evolution of GSM) and 3G technologies, a new packet-switched core network is introduced. When both the circuit-switched core network and the packet-switched core network exist, the circuit-switched network provides the mobile device connectivity with the traditional landline network such as the public switched telephone network (PSTN), and the packet-switched core network provides the mobile device connectivity with the internet and other IP networks. Compared to circuit switching, packet switching is more efficient. For example, in the case of circuit switching, a dedicated circuit is reserved between the device and the edge of the wireless network and no other mobile device can use this dedicated circuit. In contrast, a given connection can be shared among multiple users in the case of packet switching. When the user traffic arrives in bursts, which is common for applications such as email and web browsing, the dedicated resources (such as a reserved transport channel with a given bandwidth) remain unused during the data inactivity periods. Packet switching allows more efficient use of resources because the resources are not dedicated and are available to any active traffic. Therefore, since non-dedicated resources are allocated, packet-switched connections require a quality of service (QoS) profile that makes it possible to differentiate QoS aspects such as guaranteed bit rate or priority in a data connection. GPRS introduced such a QoS profile, which can be considered a policy on how to treat the forwarding of data through network entities.

GSM was developed in the late 1980s and first deployed in 1991 [5] and is one of the first 2G standards that utilized digital modulation. GSM provided several significant innovations, including:

- Digital modulation and digital data transmission (up to 14.4 kbps)
- A competitive service environment in Europe
- Worldwide interoperability
- Facilitation of low-cost wireless systems through large-scale manufacturing brought on by universal adoption

GSM spawned the beginning of data transmission service, which is now more common than voice service. GSM provided the basis for GPRS, which offers short messaging services (SMS) and enhanced throughput (up to 42.9 kbps). GPRS enhanced GSM with better data management through scheduling, and introduced a new packet-switched core network for more efficient interfacing with the internet, as mentioned earlier. GPRS is sometimes referred to as 2.5G, while EDGE (enhanced data rates for GSM evolution) is often thought of as a 2.75G technology and is also backwardly compatible with GSM.

EDGE brought reduced latency and increased throughput while using the GSM/EDGE core networking hardware and software. EDGE introduced higher-order modulation and incremental redundancy to allow for higher data rates (up to 236.8 kbps). This throughput-enhancing approach is still being improved on today to increase the throughput of modern communications [1].

GPRS and EDGE build upon and are backwardly compatible with GSM. Consequently, they are viewed as additions to 2G or evolution of 2G. GPRS allows users to utilize more than one timeslot for data and introduced different coding schemes to handle harsh as well as benign channels. While GPRS still relies on GMSK modulation, EDGE incorporated 8-PSK and more coding rate alternatives to greatly improve peak data rates. The trend of providing higher-order modulation and better radio resource management (such as scheduling) to increase throughput has continued to improve the development of 3G, 4G and 5G technologies, becoming more efficient and sophisticated with each generation. The core network developed for GPRS continued to evolve and formed the basis for the 3G and 4G core operations. Even a 5G core network has certain similarities to the original pioneering GPRS core network.

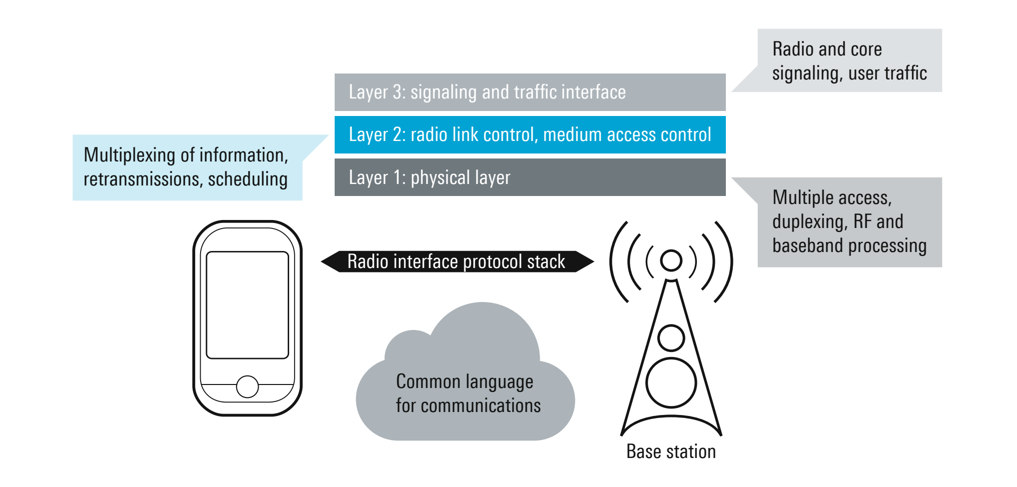
While GSM has been phased out in some countries over the years, practically disappearing in the US, it has found new life as extended coverage GSM IoT (EC-GSM-IoT), which is a low-power wide area cellular technology based on GPRS. It is well suited for IoT applications, which are relatively low in data rate, require low power (battery life of up to 10 years) but have wide coverage. Furthermore, EC-GSM-IoT can be deployed through simple software upgrades on existing GSM infrastructure.

2 Overview of radio interface processing

2.1 An overview of the radio interface protocol stack

Communications between the device and the base station occurs using the technology-specific radio interface protocol stack. The protocol stack defines a common language for air interface communications between the device and the base station. Fig. 6 shows a simplified radio interface protocol stack[6]. The exact names and processing of the layers of the radio interface protocol stack are technology-dependent.

Fig. 6: Radio interface protocol stack


Radio and core signaling, user traffic
Layer 3: signaling and traffic interface
Layer 2: radio link control, medium access control
Multiplexing of information, retransmissions, scheduling
Layer 1: physical layer
Multiple access, duplexing, RF and baseband processing
Radio interface protocol stack
Common language for communications
Base station

[6] The radio protocol stack has two or three cellular technology-specific layers. The endpoints of the communications link such as the mobile device and the server (e.g. a website server) implement the upper three layers of the five-layer TCP/IP protocol stack: layer 5, the application layer containing applications such as HTTP for web browsing; layer 4, the transport layer containing protocols such as transmission control protocol (TCP); and layer 3, the network layer containing internet protocol (IP). The TCP/IP protocol stack collapses the traditional seven-layer OSI model into a five-layer model.

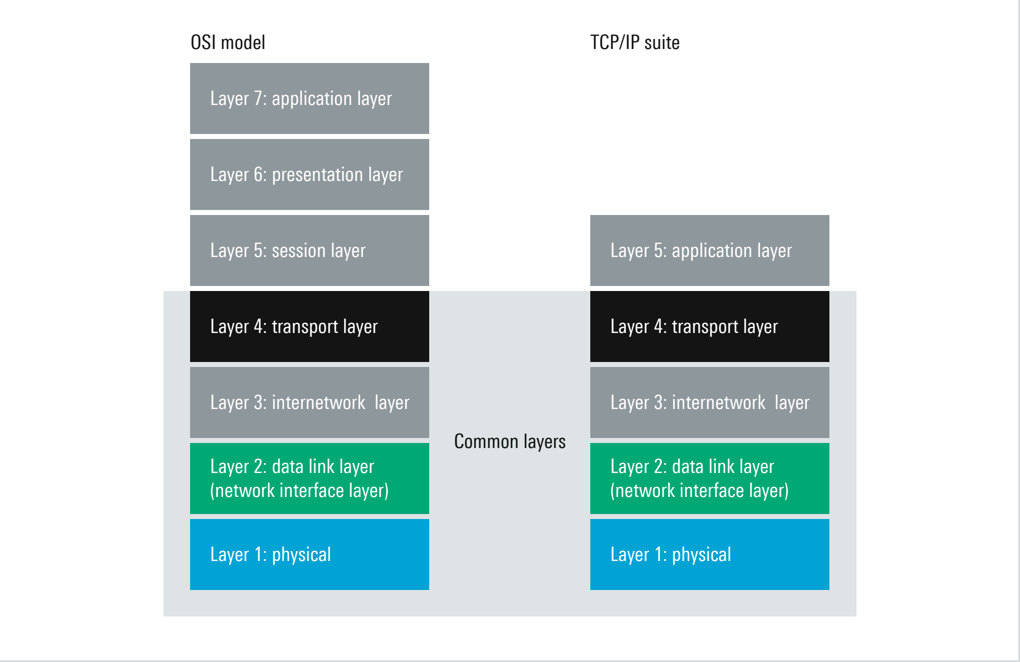
The protocol stack can be viewed as three layers (layers 1, 2 and 3). Layer 1 is the physical layer that supports features such as duplexing and multiple access as well as redundancy through adaptive channel coding and modulation. Details of the physical layer for CDMA and OFDMA respectively are provided in sections 2.2 and 2.3.

Layer 2 can have sublayers such as packet data convergence protocol (PDCP), radio link control (RLC)/radio link protocol (RLP) and medium access control (MAC). Layer 2 helps multiplex different types of information such as signaling and traffic for transmission at a given instant. The physical layer is operated at a relatively high error rate (e.g. around 1 %) to optimize the use of precious radio resources, with error correction and retransmission used to effectively lower the error rate. However, many services, including web browsing and video streaming, require a much lower error rate (e.g. 0.0001 %). That is where layer 2 comes into the picture. When the physical layer experiences an error, layer 2 retransmits a packet. Two main types of retransmissions are RLC retransmissions and MAC/physical layer retransmissions. RLC retransmissions occur relatively slowly such as on the order of tens or hundreds of milliseconds. MAC/physical layer retransmissions, often called hybrid automatic repeat request (HARQ) retransmissions, occur much faster such as on the order of few milliseconds. Since retransmissions are carried out only when needed, precious radio resources are efficiently used due to the joint work by layer 2 and layer 1. Layer 2 also carries out a critical scheduling task. Scheduling is performed by the MAC sublayer of the base station. The base station scheduling algorithm allocates physical layer radio resources to the devices for the downlink and the uplink so that signaling and traffic can be transferred over the radio interface.

Layer 3 for signaling helps exchange signaling between the device and the base station, signaling between the device and the core network. User traffic between the device and the core network typically involves transmission of IP packets using layer 1 and layer 2 of the radio protocol stack.

Let us contrast the radio interface protocol stack with the famous 7-layer open systems interconnection (OSI) model shown in Fig. 7. The end-to-end communications between two entities such as two computers or the mobile device and a web server can be represented by the OSI model. The seven layers of the OSI model include physical layer, data link layer, network (or internetwork) layer, transport layer, session layer, presentation layer and application layer. The 5-layer TCP/IP suite or IP stack shown in the figure is another popular model [1] that describes communications through the internet. The TCP/IP suite shares the lower four layers with the OSI model and consolidates the upper three layers of the OSI model into a single application layer. In the context of a mobile device communicating with a server through the internet, the wireless network provides layer 1 and layer 2. The mobile device implements all five layers of the TCP/IP protocol stack or seven layers of the OSI stack. The application layer supports applications or services and includes protocols such as hypertext transfer protocol (HTTP) for web browsing. The presentation layer transforms data into the format acceptable to applications. The session layer controls connections between two endpoints and supports full duplex, half duplex and simplex operations. Layer 4, where protocols such as the transmission control protocol (TCP) and user datagram protocol (UDP) exist, provides end-to-end connectivity. Layer 3 IP packets travel through the internet and between the mobile device and the wireless network using the help of layer 1 and layer 2.

Fig. 7: OSI model and TCP/IP suite



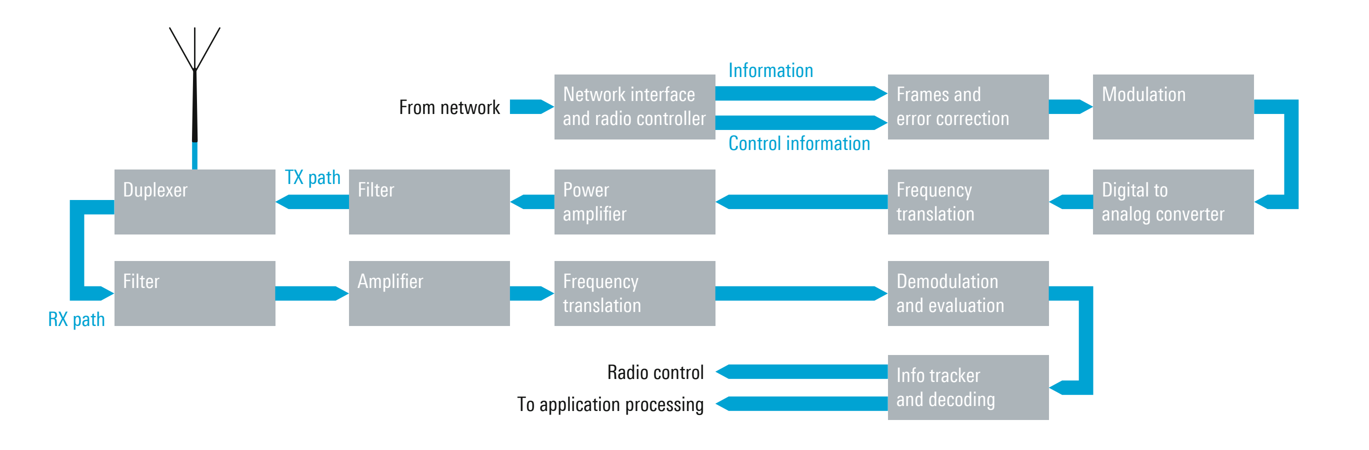
OSI model
TCP/IP suite
Layer 7: application layer
Layer 6: presentation layer
Layer 5: session layer
Layer 5: application layer
Layer 4: transport layer
Layer 4: transport layer
Layer 3: internetwork layer
Layer 3: internetwork layer
Common layers
Layer 2: data link layer (network interface layer)
Layer 2: data link layer (network interface layer)
Layer 1: physical
Layer 1: physical

Let us take an example of how user traffic such as email is sent over the air interface from the base station to the device using the radio interface protocol stack. The radio access network has received an IP packet for the device from the core network. The base station simply provides the IP packet to layer 2. Layer 2 adds suitable protocol headers to facilitate the functions of the receiver. For example, one of the layer 2 headers conveys to the receiver that the type of information is user traffic and not a signaling message. Layer 1 prepares the packet for a potentially challenging journey through the dynamic radio environment by carrying out technology-specific processing such as CDMA or OFDMA.

A closer look at the physical layer shows that the physical layer processing can be broken down into two distinct sections: a baseband section and a radio frequency (RF) section. The physical layer waveform has various properties that can impact the complexity of RF processing and how the channel will impair the signal. For instance, there is a distinct approach to the RF section of a system that uses TDD versus FDD:

- For TDD, a switch operates in realtime to separate the relatively high-power transmit signal from the low-power received signal so that the same antenna can be used for both transmission and reception
- In FDD, the transmit signal and the received signal can operate at the same time and using the same antenna, but to keep the transmit signal separate from the received signal, the signals use different frequencies and are separated using a filter called a duplexer. For both the transmitter and receiver, the key challenge is to have the same components, antennas, RF electronics and baseband processing operate over a wide range of frequency bands. An example composition of an FDD system is shown in Fig. 8.

Fig. 8: Composition of an FDD system

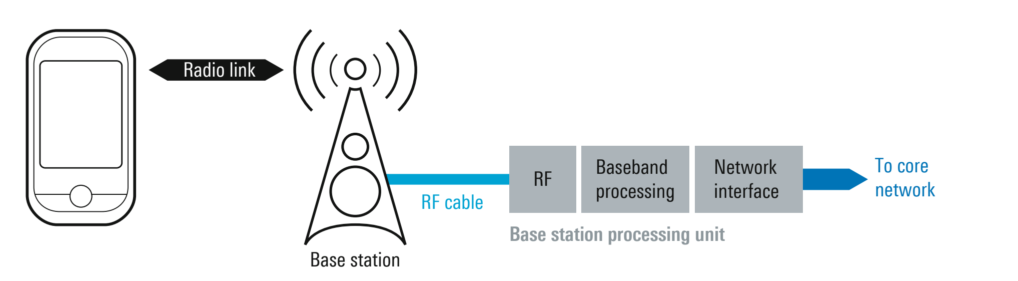

From network
Network interface and radio controller
Information
Control information
Frames and error correction
Modulation
Duplexer
TX path
Filter
Power amplifier
Frequency translation
Digital to analog converter
Filter
Amplifier
Frequency translation
Demodulation and evaluation
RX path
Radio control
To application processing
Info tracker and decoding

At the transmitter, the signal must be amplified before it is transmitted. A cellular power amplifier can be tricky to build. It must have high output power and also be able to handle significant variations in the signal power level while providing linear amplification. We quantify this capability as the peak-to-average power ratio (PAPR) of the amplifier. When multiple users are being amplified by the same power amplifier, the PAPR plays an important role in determining coverage and, in some cases, capacity. In a CDMA network, as users are added to the network, the allocation of power per user goes down since there is only so much output power available from the power amplifier in linear mode. With less power per user, the effective coverage of the CDMA network is reduced, a phenomenon that is called “cell breathing”.

Another consideration in the design of the cellular radio system is adjacent channel interference. This consideration plays a role in the design of the receiver as well as the transmitter. Out-of-band interference from the channel of a base station can impact the fidelity of communications in the adjacent channel. Consequently, effective filtering is required at the transmitter. To minimize the possibility of interference, the output level in the adjacent channel is specified by the standard. Adjacent channel interference can also impact the receiver. A receiver may have adjacent channel interference that is much stronger than the desired signal. Filtering adjacent channel signals at the RF level near the antenna has practical implementation challenges, which is why the RF circuitry at this early stage must have the ability to simultaneously handle both very weak signals and very strong signals while remaining in a linear amplification region. In other words, the receiver must have a high dynamic range. The key to achieving this is to have a good amplifier that is capable of handling a high dynamic range while contributing little noise to the signal.

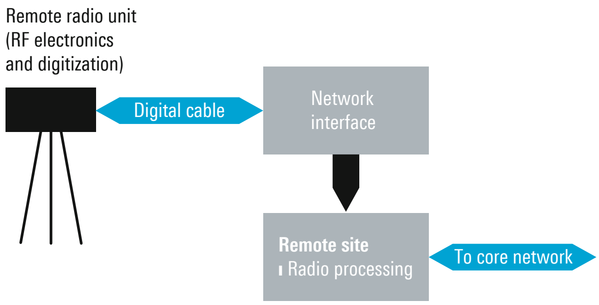
There are two approaches for deploying a cellular base station. The old approach, as shown in Fig. 9, is to link the antenna with the base station processing unit via an RF cable. This strategy has a downside in that significant losses can occur in the cables that link the antenna and processing unit. This cable loss means a loss of transmit power and an increase in received noise power.

Fig. 9: BS configuration: RF cable between the antenna and BS processing unit



The new approach is to locate the RF processing near the antenna and transport the information in digital form between the RF unit and the digital unit located at a remote site for processing as illustrated in Fig. 10. The advantage here is that the cable losses for the analog signal disappear but the downside, in some instances, is having to put the RF electronics on an antenna tower where it may be difficult to maintain. Nevertheless, this approach, which uses a remote radio head or remote radio unit, has become very popular and is cost-effective.

Fig. 10: BS configuration: digital cable between the RF unit and baseband unit



2.2 CDMA based physical layer baseband processing in 3G networks

3G cellular technologies such as UMTS, HSPA, 1x and 1x evolution-data optimized (1xEV-DO) use CDMA as the multiple access technique [1]. The key characteristic of CDMA is that the signal is spread using an orthogonal code and therefore occupies a larger bandwidth than a typical non-spread, non-CDMA signal. Section 2.2.1 describes the major physical layer baseband processing blocks of a CDMA transmitter. Section 2.2.2 describes the major physical layer baseband processing blocks of a CDMA receiver.

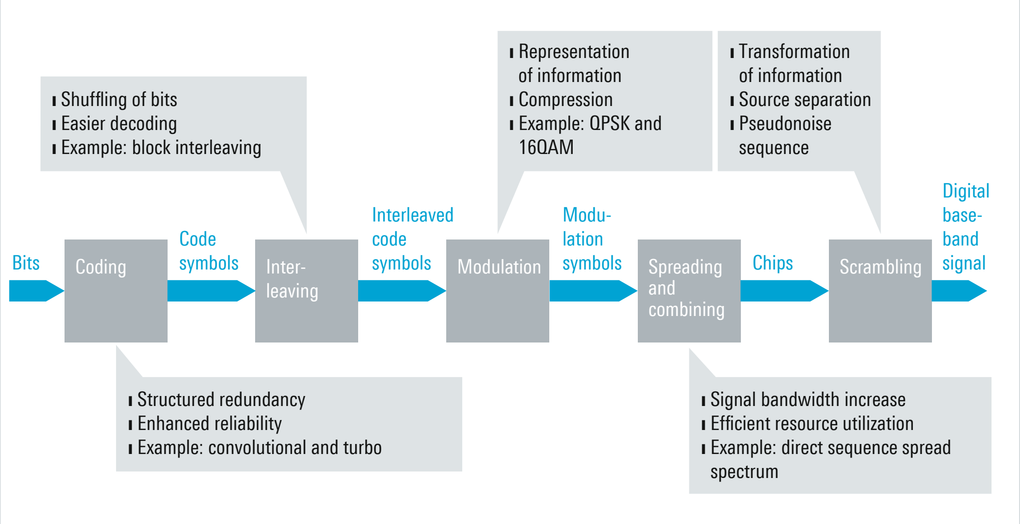
How does CDMA work?

Imagine an international party attended by citizens of different countries. Three groups of people are communicating in three different languages. Even though all the groups are talking at the same time, the members of a group are able to communicate successfully because they are talking in the same language and voices coming from other groups appear like noise. Each group can focus on its own language while ignoring the voices that are unfamiliar. That is how CDMA operates, where each language is equivalent to a code. Different codes for different users or spectrum bandwidth allow CDMA to differentiate among users within the same channel. For example, a UMTS network can simultaneously differentiate among about 128 different voice calls in 5 MHz channel bandwidth in a sector (or cell) by using different orthogonal codes for the users.

2.2.1 Physical layer baseband processing in a CDMA transmitter

Fig. 11 shows simplified baseband processing carried out in a CDMA transmitter. The overall aim of baseband processing is to reliably and efficiently transfer information over the challenging air interface.

Fig. 11: Physical layer baseband processing in a CDMA transmitter



The physical layer receives bits from the upper layer, such as user traffic plus overhead added by the radio interface protocol stack. These bits can be viewed as information bits that need to be conveyed to the receiver. The information bits are appended with cyclic redundancy check (CRC) bits so that the receiver can detect if it has correctly received the bits or not. The information bits together with the CRC bits pass through channel coding or forward error correction (FEC). Channel coding adds redundancy in a structured manner to improve transmission reliability. The simplest form of channel coding is repetition, where the same information bit is repeated multiple times. In this case, even if the radio environment corrupts some of the repeated bits, uncorrupted bits would enable the receiver to recover the original set of bits. Compared to simple repetition, coding techniques such as convolutional coding and turbo coding provide better error performance and achieve a very low error rate such as just one error out of thousands or even a million bits. Convolutional coding is computationally less intensive, making it more suitable than turbo coding when decoding needs to be performed quickly. Convolutional coding is attractive for small payloads (e.g. tens or hundreds of bits) and services such as voice. Turbo coding requires more processing power and time but has better error performance than convolutional coding, especially when the payload is relatively large. Indeed, turbo coding implementations often include two convolutional coders. Turbo coding is an appropriate choice for large payloads (e.g. thousands or more bits) and data services such as email and web browsing. In the interest of design simplification, turbo coding is also used for small payloads. For example, LTE uses turbo coding even for voice services.

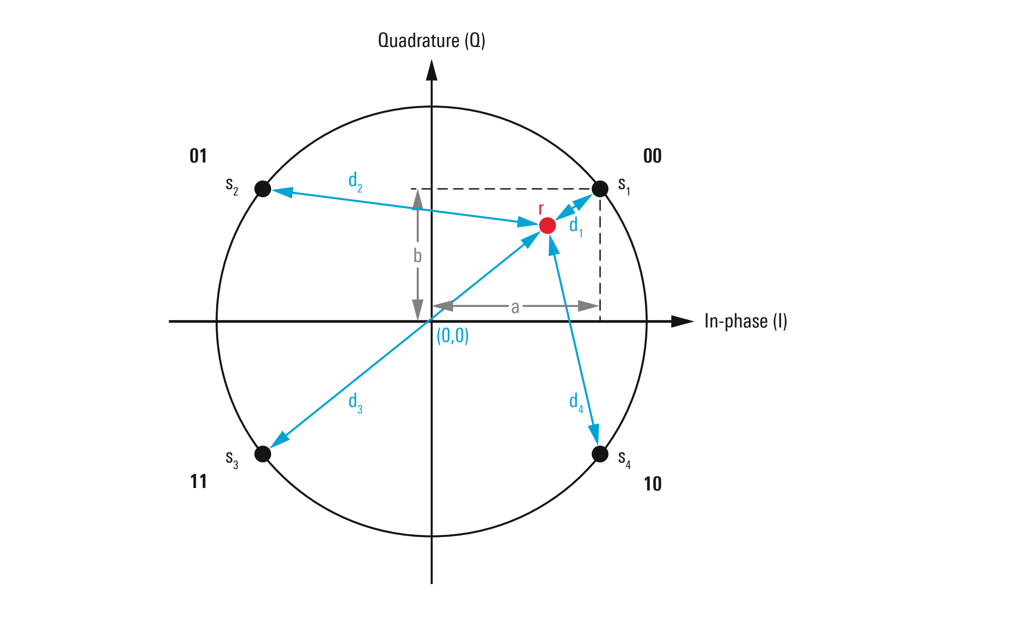
Just like shuffling playing cards changes the order of the cards, interleaving at the transmitter changes the order of code symbols. Block interleaving maps a set of input code symbols into a sequence of output code symbols that have a different order. The radio environment causes errors in consecutive code symbols. The reverse processing of interleaving, deinterleaving, is carried out at the receiver, which scatters such consecutive errors by reshuffling the bits. The decoder at the receiver can correct more errors when errors are distributed. Thus, interleaving at the transmitter and deinterleaving at the receiver facilitates decoding at the receiver.

Interleaved code symbols are digitally modulated, where a set of input code symbols is represented by a single modulation symbol. Examples of modulation schemes include binary phase shift keying (BPSK), quadrature phase shift keying (QPSK), 8-ary phase shift keying (8PSK) and 16-ary quadrature amplitude modulation (16QAM). A BPSK modulation symbol represents one code symbol, a QPSK modulation symbol represents two code symbols, an 8-PSK modulation symbol represents 3 code symbols and a 16QAM modulation symbol represents 4 code symbols. Since a modulation symbol typically represents multiple code symbols, it can be viewed as information compression and helps increase the effective data rate or throughput. Indeed, adaptive modulation and coding (AMC) involves finding the most suitable combination of the modulation scheme and the amount of redundancy introduced by channel coding for the prevailing radio channel conditions.

Example modulation and how errors occur

Consider the QPSK signal constellation diagram depicted in Fig. 12 [1]. The QPSK modulation symbol s_1 is represented by the complex number (a + j b).

Fig. 12: QPSK signal constellation



In Fig. 12, s_1 represents two information bits or code symbols 00. The other three QPSK modulation symbols are s_2, s_3 and s_4, representing the information bits 01, 11 and 10, respectively. The standards bodies such as 3GPP specify exactly the amplitudes and phases of modulation symbols for different modulation schemes. While the transmitter sends s_1, the radio channel affects the transmission such that the receiver gets the complex number r. The demodulator calculates the Euclidean distance between the received modulation symbol and all possible modulation symbols, denoted d_1 to d_4. The receiver then predicts the transmitted modulation symbol to be the one corresponding to the minimum distance, which is s_1 in Fig. 12. As the modulation order increases from QPSK to 16QAM to 64QAM and so on, the distance between the modulation symbols decreases, increasing the likelihood of the receiver making an error in estimating the modulation symbol.

Modulation symbols are multiplied by an orthogonal spreading code to increase the signal bandwidth. This is the fundamental characteristic that separates CDMA from other multiple access techniques. The chip rate after spreading is 1.2288 megachips per second (Mcps) in 3G 1x systems and 3.84 Mcps in 3G UMTS networks. The orthogonal codes are referred to as Walsh codes or Walsh-Hadamard codes in 1x networks and orthogonal variable spreading factor (OVSF) codes in UMTS networks. The chip rate is kept constant to keep the baseband signal within the target channel bandwidth, which is 1.25 MHz for 1x and 5 MHz in UMTS. Therefore, a lower data rate results in a larger spreading factor and a higher data rate results in a smaller spreading factor. For example, if the modulation symbol rate is 60k (modulation) symbols per second, a spreading factor of 64 is used to achieve the chip rate of 3.84 Mcps (60 ksps × 64 = 3840 kcps or 3.84 Mcps) for UMTS networks. And if the modulation symbol rate is 480k (modulation) symbols per second, a spreading factor of 8 is used to achieve the chip rate of 3.84 Mcps (480 ksps × 8 = 3840 kcps or 3.84 Mcps) for UMTS networks.

In the downlink, the base station allocates different orthogonal spreading codes to different users to ensure minimal (ideally zero) interference among users in the cell or sector[7]. The chips for different users are added to create a combined signal that has information for multiple users. Although the combined signal appears much like noise, there are orthogonal signals buried in this noise-like signal. In the uplink, the device can use different spreading codes for different channels so that the base station can separate out all these channels without any interference among the channels. The chips for different channels are combined.

Types of spread spectrum systems

There are two main types of spread spectrum systems: direct sequence spread spectrum (DSSS) and frequency hopping spread spectrum (FHSS). The DSSS system uses codes to spread the signal. The signal bandwidth is enlarged after spreading is applied to the data. 1x and UMTS are examples of DSSS systems. The FHSS system involves relatively narrow signal bandwidths, but the signal rapidly hops from one part of the spectrum to another part of the spectrum using a predefined pseudorandom sequence. Over a sufficiently long observation period, the signal effectively uses a much wider bandwidth than that required by the data. Since the signal uses different parts of the spectrum, the resulting frequency diversity enhances the reliability of the link. Some Bluetooth® systems and GSM use FHSS.

The combined chips coming out of the combining block are scrambled using the identity of the source of the transmission. In the downlink, the base station scrambles the chip sequence using a pseudorandom sequence that is a function of the identity of the sector (or cell). Such source-specific scrambling enables the device to distinguish the signal of one cell from the signals of other cells in the downlink. In the uplink, the device scrambles the chip sequence using a pseudorandom sequence that is a function of the identity of the device. Such device-specific scrambling in the uplink allows the base station to separate the signal of one device from the signals of other devices in the cell. The scrambled chips represent the digital baseband signal that undergoes RF processing before transmission.

2.2.2 Physical layer baseband processing in a CDMA receiver

The processing done by the transmitter prepares the signal to meet the challenges of the dynamic radio environment.

Challenges of the radio channel

What types of radio channel impairments affect transmissions? The influence of the radio environment on the signal received at the receiver can be modeled as the combination of distance-based path loss, large-scale fading and small-scale fading. The distance-based path loss reflects factors such as the distance between the transmitter and the receiver, the carrier frequency, the base station and mobile station antenna heights, and the overall environment (e.g. rural versus urban). The large-scale fading reflects the type of clutter such as buildings and vegetation. The small-scale fading, often modeled as Rayleigh and Ricean fading, reflects the influence of multipath signals and relative mobility. Multipath signals result from reflections of the transmitted signal, and CDMA systems take advantage of these multipath signals.

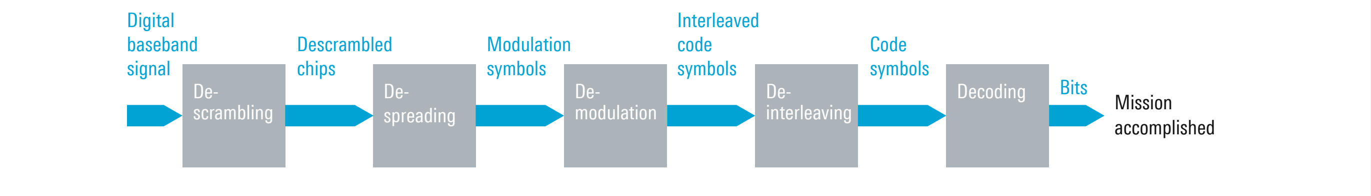
The receiver performs processing that is opposite to the transmitter processing. The receiver is much more complex than the transmitter because of the sophistication necessary to synchronize and extract the correct information as well as deal with channel impairments.

[7] In a popular cellular network configuration, one base station controls three geographic regions called sectors (or cells), where each sector covers 120° region. Three sectors together provide 3 × 120° = 360° coverage around the base station.

The RF processor at the receiver receives the RF signal from the radio interface using one or more receive antennas. After passing through stages such as filtering, low-noise amplification, frequency downconversion and analog-to-digital (A/D) conversion, the RF signal is transformed to a digital baseband signal that is ready to be processed by the baseband processor.

Fig. 13 shows how a simplified CDMA baseband receiver carries out physical layer processing. The basic job of the receiver is to undo any damage done by the radio environment and retrieve the information bits intended by the transmitter.

Fig. 13: Physical layer baseband processing in a CDMA receiver



The digital baseband signal consists of chips from the desired signal as well as any interfering signals. The first step is to extract the information from the desired source. This is done by descrambling, which is the opposite of scrambling. In the downlink, the receiver at the device correlates the incoming signal with the scrambling sequence associated with the known scrambling ID of its serving cell or sector. In the uplink, the receiver at the base station individually correlates the incoming signal with the scrambling sequences associated with the known scrambling IDs of its users. The descrambled chips of the signal are despread using suitable orthogonal codes. In the downlink, the device uses known orthogonal codes to determine the modulation symbols of interest. In the uplink, the base station uses orthogonal codes to create modulation symbols of the device's control or data channels.

The modulation symbols are demodulated to determine the interleaved code symbols. Demodulation can be coherent or non-coherent. Coherent demodulation gets the help of reference signals with known bits or signals to quantify radio channel conditions and uses such channel knowledge to clean up the modulation symbols of the channel of interest (e.g. traffic channel carrying information bits) before estimating the bits (the interleaved code symbols) represented by modulation symbols.

The interleaved code symbols from the demodulator are deinterleaved. Deinterleaving is the opposite of interleaving. Deinterleaving spreads out consecutive errors caused by the radio environment as discussed above. Such non-consecutive errors are somewhat easier for the decoder to rectify.

The deinterleaved code symbols are decoded. To decode convolutional code symbols, Viterbi decoding or its variation is widely used. To decode turbo code symbols, the maximum a posteriori (MAP) algorithm or its variation is quite popular. A turbo decoder is more complex than a convolutional decoder and requires significantly more processing power. The decoder produces the information bits, and the CRC check helps detect the presence of any error(s) in the decoded information.

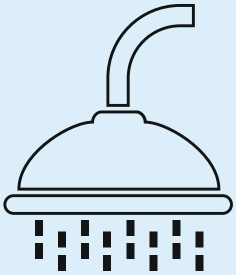
While standards bodies such as 3GPP specify virtually all aspects of the transmitter, receiver implementations tend to be implementation-specific and proprietary. The standards, however, specify minimum performance requirements for receivers (as well as transmitters) so that a certain implementation quality can be ensured and the transmitter and the receiver can communicate seamlessly. The receiver implementation is one of the important product differentiators for device and base station vendors.

Rake receiver in CDMA systems

In cellular networks, when the transmitter sends out a signal, the receiver may receive multiple copies of the same signal due to reflections. These delayed copies are called multipath signals. In non-CDMA networks, the receiver must overcome such multipath signal interference by, for example, inserting a guard time. However, CDMA systems use a special receiver called a rake receiver that can make use of multipath signals by combining the energy of those signals. The receiver's name comes from the fact that it resembles a garden rake, where each finger of the rake receiver can track a multipath signal. The multipath signals can be from one cell or sector or multiple cells or sectors. It is quite common for a rake receiver to have four or more fingers. The receiver can distinguish between two multipath signals if they arrive at least one chip period apart. Typical time difference of arrivals for multipath signals is on the order of few microseconds in a cellular network. 1x and UMTS networks have chip periods of about 0.3 μs and 0.8 μs, respectively. Therefore, rake receivers work very well in the presence of multipath signals. The rake receiver combines the energy of modulation symbols using a technique called maximal ratio combining. More reliable modulation symbols lead to better decoder error performance.

2.3 OFDM baseband processing in 4G and 5G networks

4G LTE and 5G technologies use OFDMA as the multiple access technique. The key characteristic of OFDMA is that a wideband signal is created using multiple narrowband channels called subcarriers. These subcarriers are orthogonal to one another. Section 2.3.1 describes the major physical layer baseband processing blocks of an OFDMA transmitter. Section 2.3.2 describes the major physical layer baseband processing blocks of an OFDMA receiver.



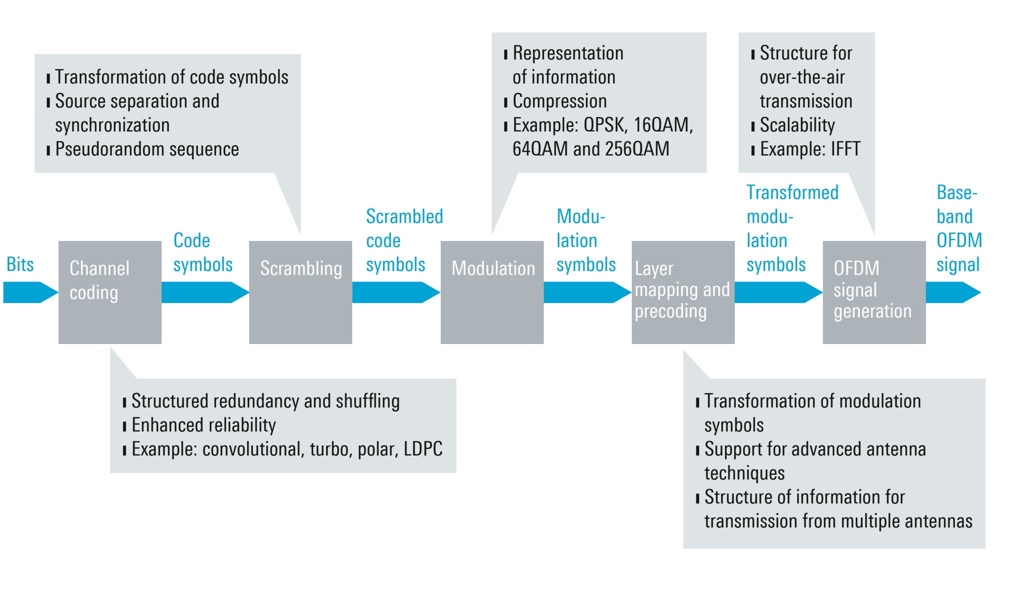
How does OFDMA work?

Imagine a showerhead distributing parallel streams of warm water. That is what OFDMA does, except OFDMA delivers data streams instead of water streams. Just like the showerhead distributes a large amount of water using multiple tiny streams in parallel, OFDMA distributes large amounts of data for different users by using different sets of narrowband frequency channels in parallel. These frequency channels are called subcarriers. The subcarriers are orthogonal to one another so there is no interference (minimal interference in practice) between the subcarriers. The available channel bandwidth is divided among multiple subcarriers. For example, in LTE, 600 subcarriers with 15 kHz subcarrier spacing are available in a 10 MHz channel bandwidth with 9 MHz transmission bandwidth (600 × 15 kHz = 9 MHz) and 1 MHz total guard band within the channel bandwidth. A group of 12 consecutive subcarriers constitutes a resource block. Users are allocated one or more resource blocks for data transmission.

2.3.1 Physical layer baseband processing in an OFDMA transmitter

Fig. 14 shows simplified baseband processing carried out in an OFDMA transmitter. Several processing blocks are common to both an OFDMA transmitter and a CDMA transmitter.

Fig. 14: Physical layer baseband processing in an OFDMA transmitter


Bits
Channel coding
Code symbols
Scrambling
Scrambled code symbols
Modulation
Modu-lation symbols
Layer mapping and precoding
Transformed modu-lation symbols
OFDM signal generation
Base-band OFDM signal
Transformation of code symbols
Source separation and synchronization
Pseudorandom sequence
Representation of information
Compression
Example: QPSK, 16QAM, 64QAM and 256QAM
Structure for over-the-air transmission
Scalability
Example: IFFT
Structured redundancy and shuffling
Enhanced reliability
Example: convolutional, turbo, polar, LDPC
Transformation of modulation symbols
Support for advanced antenna techniques
Structure of information for transmission from multiple antennas

An upper layer provides the physical layer information bits that need to be conveyed to the receiver. The channel coding block adds CRC bits, forward error correction (FEC) coding and interleaving. The CRC bits are added to the block of information bits (often referred to as a transport block) so that the receiver can detect the presence of any errors. Like 3G 1x and UMTS, LTE uses FEC techniques such as convolutional coding and turbo coding. 5G introduces FEC techniques such as polar coding and low density parity check (LDPC) coding. Polar coding is used as a replacement for convolutional coding to protect certain types of signaling messages such as resource allocation signaling messages from the base station to the device and a channel quality report from the device to the base station. Polar coding is chosen instead of convolutional coding because it is more reliable than convolutional coding. Polar coding can achieve a lower error rate than convolutional coding for a given radio channel condition. LDPC coding can achieve a lower error rate than turbo coding and consumes less processing power (and consequently battery power) than turbo coding, especially for large packets, which is why LDPC coding is selected in 5G instead of turbo coding. Like in a CDMA transmitter, interleaving is also used in an OFDMA transmitter to reorder the sequence of code symbols to facilitate decoding at the receiver.

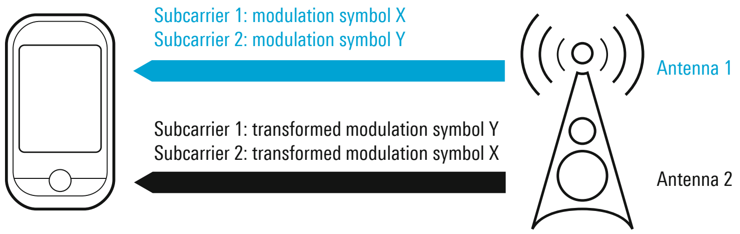
Interleaved code symbols are scrambled using a pseudorandom sequence. Such scrambling helps separate one communications link from another. For example, in the LTE downlink, the pseudorandom sequence used on a channel carrying user traffic is a function of the identity of the cell (e.g. physical cell identity or PCI in LTE and 5G), the radio interface identity of the device (e.g. cell radio network temporary identifier (C-RNTI) allocated to the device by the base station for use in a sector) and the identity of a timing unit (e.g. slot number “1” where a slot occupies a 0.5 ms interval). The receiver can separate out its own signal from all the other users’ signals in the same resources, such as the same subcarriers allocated to users in different sectors.

Interleaved and scrambled code symbols are digitally modulated, where a set of input code symbols is represented by a single modulation symbol. Higher-order modulation schemes are available for use in 4G and 5G OFDMA systems than in 3G networks such as 1x and UMTS. Typical modulation schemes supported for the traffic-carrying channels include QPSK, 16QAM, 64QAM and 256QAM. A QPSK modulation symbol represents 2 code symbols, a 16QAM modulation symbol represents 4 code symbols, a 64QAM modulation symbol represents 6 code symbols and a 256QAM modulation symbol represents 8 code symbols. Higher-order schemes such as 64QAM and 256QAM significantly increase the effective data rate or throughput when the radio channel conditions correspond to a high signal-to-interference-plus-noise ratio (SINR). AMC is used in OFDMA just like in CDMA. Since the OFDMA time structures in 4G LTE and 5G allow short transmission time intervals compared to 3G, more opportunities for fast channel-sensitive AMC exist in OFDMA systems.

Modulation symbols undergo layer mapping and precoding, which facilitate implementation of advanced antenna techniques such as transmit diversity[8] and spatial multiplexing. These antenna techniques make use of two or more transmit antennas and/or two or more receive antennas. Transmit diversity involves transmitting the same information (e.g. a given modulation symbol) on different dimensions. Examples of basic dimensions are space, time and frequency. Due to the dynamic nature of the radio environment, the probability of multiple dimensions simultaneously experiencing poor radio environment is quite low. So if the propagation path associated with one dimension (e.g. an antenna or a subcarrier) experiences a challenging radio environment such as a signal fade at a given instant, an additional propagation path associated with another dimension can still carry the modulation symbol successfully. Fundamental transmit diversity techniques are space diversity, time diversity and frequency diversity. Space diversity sends the same information using different spaces or antennas. Time diversity involves sending of the same information at different times. Frequency diversity implies that the same information is sent using different frequencies or subcarriers.

LTE uses the transmit diversity scheme called space frequency block coding (SFBC), which combines the benefits of both space diversity and frequency diversity. As shown in Fig. 15, SFBC involves transmitting the same modulation symbol on two different antennas using two different subcarriers. For example, from antenna 1, modulation symbol X is sent using subcarrier 1, and modulation symbol Y is sent using subcarrier 2. From antenna 2, transformed modulation symbol X is sent using subcarrier 2, and transformed modulation symbol Y is sent using subcarrier 1[9]. Since two different antennas are used to send the same modulation symbol, the benefits of space diversity are realized. Similarly, since two different subcarriers are used to transmit the same modulation symbol, frequency diversity is utilized. The exact forms of the modulation symbols sent from the transmit antennas are different (e.g. suitable phase shifts) to facilitate the receiver's job of combining the received signals to retrieve the original modulation symbols. SFBC improves the reliability of communications through redundant transmissions in different space and frequency dimensions. SFBC can be used even when the radio channel conditions are hostile, such as low SINR situations near the cell edge.

Fig. 15: Space frequency block coding: transmit diversity in LTE

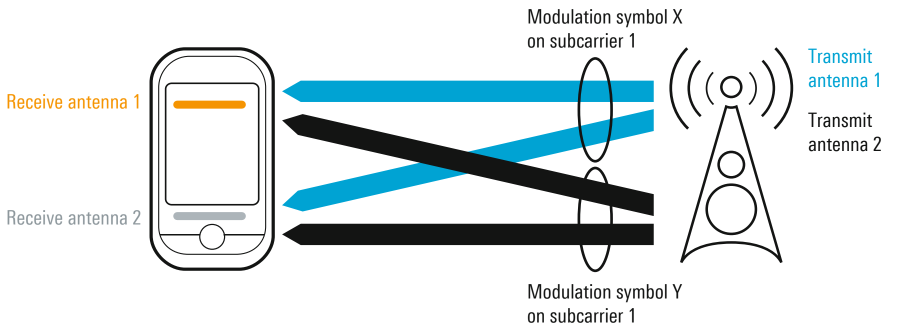

Subcarrier 1: modulation symbol X
Subcarrier 2: modulation symbol Y
Antenna 1
Subcarrier 1: transformed modulation symbol Y
Subcarrier 2: transformed modulation symbol X
Antenna 2

[8] Receive diversity involves reception of signals on two or more antennas. Receive diversity can be (and often is) implementation-specific and does not need any explicit standardization requirements.

[9] A modulation symbol can be transformed by using a different phase shift relative to the original modulation symbol.

While transmit diversity aims to enhance reliability through redundancy, spatial multiplexing aims to increase data rate or throughput. Two main spatial multiplexing techniques supported in LTE are single-user multiple input multiple output (SU-MIMO) and multi-user multiple input multiple output (MU-MIMO). Digital beamforming is supported as a special case of MIMO. SU-MIMO involves the simultaneous transmission of multiple spatial layers for a given user, where different layers transmit distinct modulation symbols using the same radio resources but with the help of multiple antennas. For example, consider 2x2 SU-MIMO[10] transmission in the downlink shown in Fig. 16. Transmit antenna 1 sends modulation symbol X from a subcarrier such as subcarrier 1. Furthermore, transmit antenna 2 sends modulation symbol Y from the same subcarrier 1. Both receive antenna 1 and receive antenna 2 receive the signal transmitted by transmit antenna 1. Note that the radio channel may affect the transmitted modulation symbol X differently because the two propagation paths are different: "transmit antenna 1 to receive antenna 1" path and "transmit antenna 1 to receive antenna 2" path. Similarly, transmit antenna 2 sends modulation symbol X from the same subcarrier 1. Both receive antenna 1 and receive antenna 2 receive the signal transmitted by transmit antenna 2. The radio channel may also affect the transmitted modulation symbol Y differently on two different propagation paths, "transmit antenna 2 to receive antenna 1" path and "transmit antenna 2 to receive antenna 2" path. Since the same subcarrier is used on both transmit antennas to send two different modulation symbols, the date rate doubles compared to single-antenna transmission and single-antenna reception. Furthermore, the receive antenna 1 signal is some combination of modulation symbols X and Y such as Rx1 = X+Y, and the receive antenna 2 signal is some combination of modulation symbols X and Y such as Rx2 = X-Y depending on the radio environment. The job of the receiver is to process Rx1 and Rx2 to recover the original modulation symbols X and Y. The receiver estimates so-called spatial signatures by observing how the radio channel influences the reference signals and uses such spatial signatures to determine exactly how to process Rx1 and Rx2 to retrieve the modulation symbols. Since the use of two antennas implies space or spatial multiplexing, two spatial multiplexing layers are said to be used.

Fig. 16: Single-user multiple input multiple output (SU-MIMO): spatial multiplexing in LTE

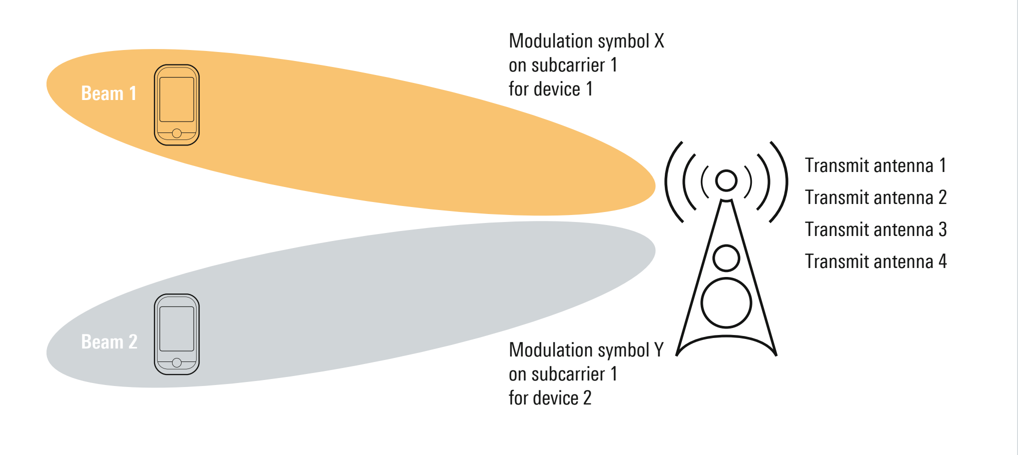


Multi-layer transmission in SU-MIMO requires the receiver to be able to distinguish among the transmit signals. In some channel conditions, the receiver may not be able to separate out the transmit signals coming from different transmit antennas. In such cases, multi-layer transmission is not feasible but digital beamforming is possible, where suitable phase adjustments for a given modulation symbol for different transmit antenna signals result in more cumulative energy at the receiver. The phase adjustments allow the different signals from each of the transmitter's antennas to add coherently at the position of the receiver's antenna. While SU-MIMO involves the transmission of multiple layers, digital beamforming is often a single-layer transmission case.

[10] SU-MIMO is an MxN downlink, where M indicates the number of transmit antennas in a sector at the base station and N indicates the number of receive antennas at the device.

MU-MIMO involves the simultaneous transmission of multiple spatial layers, where different layers transmit distinct modulation symbols using the same radio resources for different users. Consider an example of MU-MIMO downlink transmission shown in Fig. 17. The use of more transmit antennas enhances beamforming in support of MU-MIMO.

Fig. 17: Multi-user multiple input multiple output (MU-MIMO): spatial multiplexing in LTE



There are various ways to do MU-MIMO. For example, MU-MIMO in the downlink, the base station can send different modulation symbols to different users at the same time and for the same subcarrier. In one possible simplistic configuration, as shown in Fig. 17, one beam is created for one user to send a modulation symbol X on subcarrier 1. The base station can create a separate beam for another user and then send modulation symbol Y on subcarrier 1 using a second beam. These two beams reuse the same radio resources in a given sector. In general, information going to antennas can be weighted such that device 1 experiences relatively larger power for its own information while minimal interference is caused to other devices. Likewise, device 1 is receiving minimal interference from the signal intended for device 2 because of the appropriate weighting applied to antenna signals. This is possible because MU-MIMO exploits the fact that the radio channels are different for spatially-separated devices. Since the users are separated in the space domain via spatial beams, MU-MIMO is a general extension of space division multiple access (SDMA). The device receives the signals on one or more receive antennas to retrieve its modulation symbols. MU-MIMO increases the number of users that benefit from high data rates due to the reuse of radio resources (such as frequencies) in the sector.

Note that in the early days of MIMO research, MIMO was promoted erroneously as a way to beat the Shannon-Hartley law. This is not the case; what makes MIMO high capacity is that it uses different spatial channels for conveying the information. Information is sent from different transmit antennas and received on multiple receive antennas with multiple transmit-receive propagation paths.

Consider the use of layer mapping and precoding to implement SFBC when two transmit antennas in a sector are used at the base station. For SFBC, layer mapping and precoding would result in the placement of two modulation symbols X and Y in specific forms such as their original structures and transformed structures[11)] on two antennas and two consecutive subcarriers as shown in Fig. 15. The receiver is informed about the use of SFBC and can process the signals to combine the signals received on different subcarriers in such a manner that the original modulation symbols X and Y can be retrieved.

[11)] If the modulation symbol is X = a + jb, its transformed structure would be the complex conjugate X*, which is a – jb. Transmissions of the original and transformed modulation symbols make it easier for the receiver to extract X and Y. For example, one antenna transmits X on subcarrier f_1 and Y on subcarrier f_2, while the second antenna transmits Y* on f_1 and –X* on f_2.

Now, consider the use of layer mapping and precoding to implement SU-MIMO when two transmit antennas in a sector are used at the base station. For 2x2 SU-MIMO in the downlink, layer mapping and precoding can place two modulation symbols X and Y in specific forms (e.g. original structures or potentially transformed structures[12]) on two antennas on the same subcarrier as shown in Fig. 16. The receiver is informed about the MIMO transmission parameters and can process the signals to retrieve the original modulation symbols X and Y.

Modulation symbols in suitable forms for transmission on specific antenna ports are given to the OFDM signal processing block. A distinct OFDM signal is generated for each transmit antenna. Inverse fast Fourier transform (IFFT) combines subcarriers carrying layer-mapped and precoded modulation symbols and creates an OFDM symbol consisting of a time-domain series of samples. The OFDM symbol is a combination of the modulation symbols carried on all subcarriers. Since OFDMA is used in the downlink, such a symbol is often called an OFDMA symbol. When the LTE channel bandwidth is 10 MHz, a 1024-point IFFT is used to generate a series of 1024 samples during the symbol period of 66.6 µs by combining modulation symbols on six hundred subcarriers. The subcarriers are spaced 15 kHz apart, leading to a useful symbol period of 1/15 kHz = 66.6 µs. While CDMA transmitters send out chips over the air, OFDM transmitters send out OFDM symbols over the air. If there are four transmit antennas, four distinct baseband OFDM symbols are generated with one symbol per transmit antenna. The digital baseband signals in the form of OFDM symbols are sent over the air after undergoing RF processing.

SC-FDMA: OFDMA with a twist

LTE uses single-carrier frequency division multiple access (SC-FDMA) instead of pure OFDMA in the uplink. At a given instant, the OFDM/OFDMA sample involves contributions from modulation symbols carried by multiple subcarriers. This combining of essentially random numbers leads to high PAPR for OFDM/OFDMA and therefore the cell-edge devices have relatively less average power available for communications when OFDM/OFDMA is used in the uplink. Such transmit power constraints can adversely affect reliability and/or uplink throughput. LTE uses a variation of OFDMA in the uplink called SC-FDMA. SC-FDMA can reduce PAPR by 3 dB or even more. At the transmitter, SC-FDMA involves passing the modulation symbols through discrete Fourier transform (DFT) and the DFT output symbols are placed on subcarriers by IFFT. Since DFT and IFFT are inverse operations, the result is that the SC-FDMA sample at any instant is influenced by one modulation symbol rather than numerous modulation symbols. The absence of combining multiple modulation symbols at a given instant enables SC-FDMA to reduce PAPR compared to OFDMA, improving cell-edge reliability and/or uplink throughput while reducing power consumption.

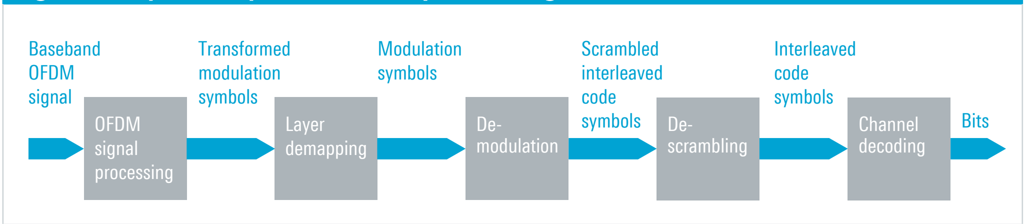
2.3.2 Physical layer baseband processing in an OFDMA receiver

The RF processor at the OFDMA receiver receives the RF signal from the radio interface using one or more receive antennas. After passing through stages such as filtering, low-noise amplification, frequency downconversion and A/D conversion, the RF signal acquired from a given receive antenna is transformed into a digital baseband physical layer signal that is ready to be processed by the baseband processor.

Fig. 18 shows how a simplified OFDMA baseband receiver carries out physical layer processing. As mentioned earlier, receiver implementations tend to be implementation-specific and proprietary. Of course, 3GPP specifies many minimum performance requirements for receivers to ensure a certain implementation quality.

[12] In one configuration, the base station may send X from one antenna and Y from another antenna using the same subcarrier. In another configuration, the base station may send X + Y from one transmit antenna and X – Y from a second transmit antenna using the same subcarrier. See 3GPP TS 36.211 for all possible supported configurations.

Fig. 18: Physical layer baseband processing in an OFDMA receiver



The digital baseband signal is received for each receive antenna. It is common for an OFDMA receiver to have two or four receive antennas. A suitable number of samples of the received OFDM signal on a given receive antenna is taken during the total OFDMA symbol time based on the value of cyclic prefix (CP), a set of symbols that resides both at the front and the back of the OFDMA symbol. The overlap of multipath signals occurs in the region of the CP. The redundancy allows the information to be extracted. In other words, intersymbol interference (ISI) might have corrupted the initial portion of the OFDMA symbol. If the CP is adequate for a given radio environment, the useful symbol period of 66.6 µs would not have any degradation due to ISI.

OFDM signal processing, shown in Fig. 18, includes processing by a fast Fourier transform (FFT). The FFT uses a block of 1024 samples over the symbol period of 66.6 µs in the case of 10 MHz channel bandwidth and 600 subcarriers. The FFT is used to retrieve transformed modulation symbols present on each of the subcarriers.

Layer demapping considers the layer mapping used by the transmitter and conveyed to the receiver to recover original modulation symbols. For example, modulation symbols might have been transformed into different structures due to layer mapping and precoding at the transmitter.

The goal of demodulation is to estimate scrambled interleaved code symbols based on the modulation symbols. The exact processing in demodulation depends on the type of antenna technique used by the transmitter, e.g. transmit diversity or spatial multiplexing. To help the receiver carry out coherent demodulation reliably, the transmitter transmits orthogonal reference signals. The receiver can estimate radio channel conditions based on such reference signals and use the knowledge to improve the estimates of scrambled interleaved code symbols.

Descrambling uses the known scrambling sequence to descramble the signal and estimate interleaved code symbols. Recall that the scrambling sequence is a function of the cell identity and the device identity on the radio interface for the shared channel carrying user traffic.

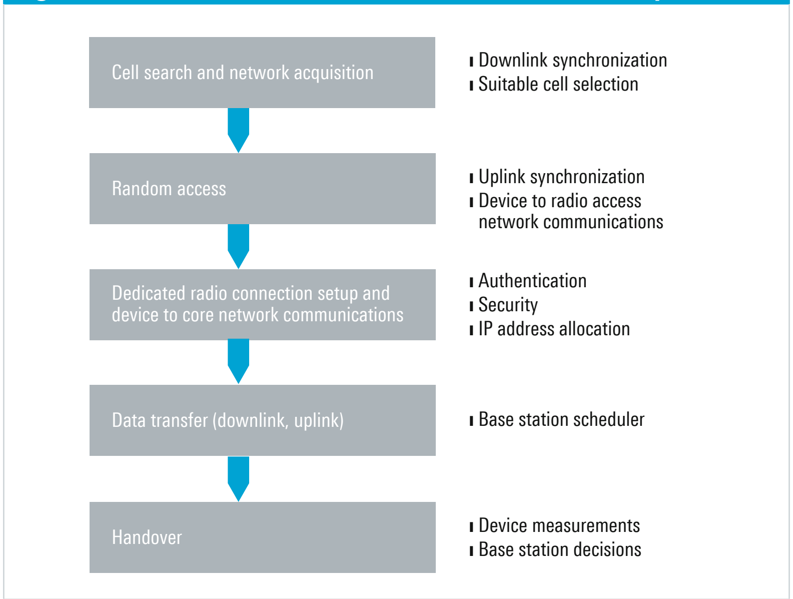
Interleaved code symbols are processed by channel decoding, where deinterleaving is carried out first, followed by decoding of code symbols such as those generated by convolutional coding, turbo coding, polar coding or LDPC coding to produce the desired information for that user. Decoding of convolutional coding and turbo coding is similar for CDMA and OFDMA. Polar decoding and LDPC decoding are quite different. To decode polar code symbols, a successive cancellation decoder or its variants can be used. To decode LDPC code symbols, a message-passing belief propagation algorithm can be used. Like other receiver functions, decoding algorithms are implementation-specific. Once decoding has been completed, the original bits are obtained and passed to the upper layers of the radio interface protocol stack.

3 Device network radio interface interactions

3.1 Overview of device operations

The radio interface interactions between the device and the radio access network are carried out with the help of the technology-specific radio interface protocol stack. Before the device can exchange any traffic such as emails and videos, certain prerequisites have to be met. Fig. 19 shows a typical series of operations from the power-up network acquisition to data transfer and handover [1].

Fig. 19: Overview of device radio access network operations



When a device is turned on, it needs to detect the cellular network and learn about the network so that it can contact the network. The device looks for specific signals broadcast by the network. Such signals help the device synchronize with the network so that it can look for relevant signals and channels to execute the remaining operations[13].
After the device learns adequate information about the network such as the identity of the network (or the wireless service provider), the device contacts the base station using common uplink resources such as the random access channel. Until dedicated resources are assigned to the device, the device and the network need to reply on common channels and signals in the downlink and uplink for communications. Now that the base station is aware of the existence of the device, a dedicated radio connection between the device and the base station can be set up. Since the authenticity of the device needs to be confirmed and since such authentication is the responsibility of the core network, a connection between the device and the core network is established. Suitable security between the device and the network is also activated. Relevant links among the network nodes, from the device to the edge of the network (e.g. a gateway facing the internet), are set up to facilitate end-to-end data transfer. The network assigns one or more IP addresses to the device since typical applications and services are IP-based. Now that all prerequisites are met, data transfer can take place. The device can move from one cell to another and one base station to another, which results in a handover. Sections 3.2 to 3.6 take a closer look at these operations.

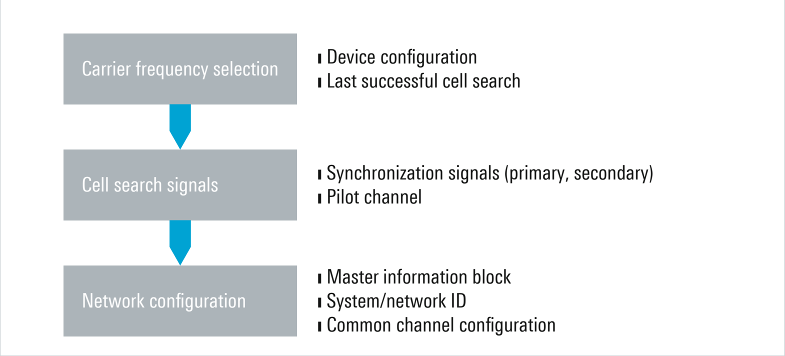
[13] Cellular standards often distinguish between a physical signal and a physical channel. A physical signal exists only at the physical layer and does not carry any upper layer information. A physical channel, on the other hand, carries information originating at upper layers of the radio interface protocol stack.

3.2 Cell search and cellular network acquisition

When a device is turned on, it searches for a suitable cell on a carrier frequency using well defined technology-specific signals. Successful cell search implies that the device has achieved time and frequency synchronization in the downlink and that the device can communicate with the found cell with satisfactory radio link quality.

Fig. 20 summarizes the overall cell search process.

Fig. 20: An overview of the cell search process


Carrier frequency selection
Device configuration
Last successful cell search
Cell search signals
Synchronization signals (primary, secondary)
Pilot channel
Network configuration
Master information block
System/network ID
Common channel configuration

The exact mechanism for the cell search is implementation-specific. However, the device is typically configured with a set of preferred carrier frequencies, such as those commonly used by a service provider. For example, an LTE device may be configured to look for a carrier frequency first in the 700 MHz band and then in the PCS band. If the device has successfully detected a cell on a certain carrier frequency in the past, storing and using the information about the detected carrier frequency accelerates the cell search.

While observing a given carrier frequency, the device looks for well-known technology-specific signals to get synchronized with the network. For example, the device looks for a pilot channel in the case of a 3G 1x network and for the primary synchronization signal and the secondary synchronization signal in the case of a 3G UMTS network, a 4G LTE network or a 5G network. The sequences of bits or symbols on the signals relevant to the cell search (such as the pilot channel and the synchronization signals) are well defined in the standards so that the base station knows what to transmit in a given cell or sector and the device knows what to look for. Since the technologies transmit signals and channels at specific times using technology-specific radio resources, time and frequency synchronization are quite critical.

After achieving synchronization in the downlink, the device continues to observe the downlink to learn about the configuration of the network. 3G technologies such as 1x and UMTS use the same channel bandwidth for cell search signals (e.g. synchronization signals) and non cell search signals/channels (e.g. channels carrying system information and user traffic). However, technologies such as LTE and 5G typically use different bandwidths for cell search signals and non cell search signals and channels. In LTE and 5G, the device learns about the bandwidth of non cell search signals by observing messages such as the master information block (MIB). The device also finds out the identity of the cellular network (or the service provider). In 3G UMTS, 4G LTE and 5G, the network identity is represented by the public land mobile network (PLMN) [14]. 3G 1x uses the system or network ID pair to identify a given operator's network.

[14] The PLMN consists of two parts, the mobile country code (MCC) and the mobile network code (MNC). PLMNs are globally unique. Different countries have their own MCCs, and cellular service providers in a given country have one or more MNCs. The system and network ID pairs used in 1x and 1xEV-DO systems are also globally unique.

The device also evaluates the need to look for another cell. For example, if a 4G or 5G network has specified a certain signal level threshold (e.g. –110 dBm) and if the device receives the cell's signal below this threshold, the device needs to shop around for a better quality cell. Otherwise, the device continues to stay with the current cell. The device obtains from the network relevant information about common channels such as the uplink random access channel. Now the device is all set to contact the network.

3.3 Contacting the cellular network: random access

So far, all the action has been in the downlink because the device has been processing signals/channels in the downlink. The base station has no idea about the device's existence yet. Fig. 21 illustrates how the device establishes contact with the radio access network.

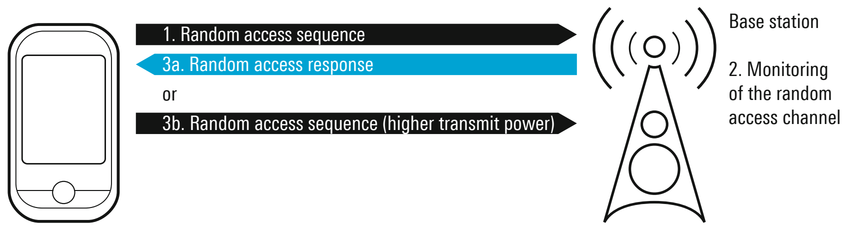
Until dedicated resources are allocated to a device, common channels or signals are used. In the uplink, the device uses the common random access channel to contact the base station. The random access channel uses well-known radio resources. The device sends a known sequence of bits or symbols on the random access channel at the transmit power level that can overcome the estimated path loss between the device and the base station. This sequence is randomly chosen from a set of sequences to minimize the probability of two devices using the same sequence.

The base station is always on the lookout for the random access sequences on the time-frequency resources designated for the random access channel. If the base station detects one or more sequences, it sends a response to the device(s) on a common downlink channel. This response is associated with the detected random access sequence.

If the device does not detect any response from the base station associated with the random access sequence, it assumes that the base station did not detect the sequence. It then randomly chooses another sequence and transmits this sequence at a time instant randomly chosen within a time window. The transmit power level of the sequence is chosen to be higher than the last transmission of the random access sequence to increase the likelihood of detection by the base station. Parameters such as the allowed power increase between successive transmissions of the sequences and the maximum number of attempts are obtained by the device during the network acquisition stage.

If the random access procedure is successful, the base station allocates dedicated radio resources to the device to ensure reliable communications. If the device remains unsuccessful in getting a response from the base station, it typically re-enters the cell acquisition stage.

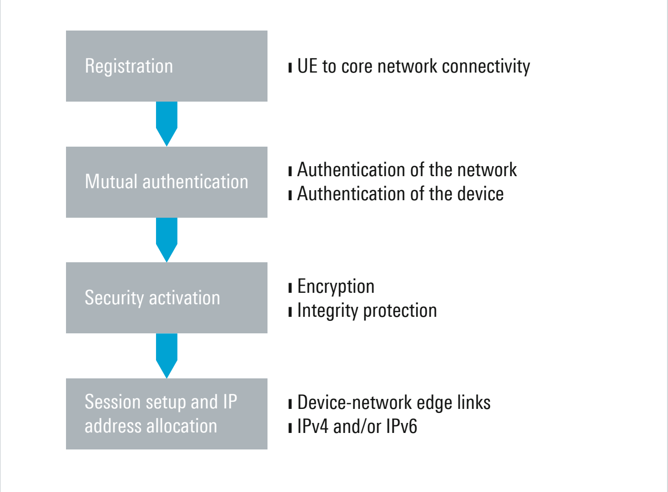
Fig. 21: An overview of the random access procedure


1. Random access sequence
3a. Random access response
or
3b. Random access sequence (higher transmit power)
Base station
2. Monitoring of the random access channel

3.4 Device core network interactions

Assigning dedicated resources to the device when a random access procedure is successful results in a dedicated radio connection between the device and the base station in a given cell or sector. This radio connection, along with the connection between the radio access network and the core network, enables the device and the core network to carry out procedures such as registration, authentication, security activation, session or bearer setup and IP address allocation as shown in Fig. 22[15].

Fig. 22: Device core network interactions


Registration
UE to core network connectivity
Mutual authentication
Authentication of the network
Authentication of the device
Security activation
Encryption
Integrity protection
Session setup and IP address allocation
Device-network edge links
IPv4 and/or IPv6

When the device registers with the core network, the core network becomes aware of the device's reachability in a given geographic area. Such registration enables the network to page the device for incoming traffic (e.g. receiving a phone call or email) when the device is in idle mode. Note that the dedicated radio connection between the device and the network is released if there is no data activity for a few seconds. When dedicated radio resources are released, the device enters the idle mode. In the idle mode, there is no dedicated radio connection between the device and the radio network.

The device relies on the common channels and signals for communication with the radio network. Without idle mode, the device would have consumed processing power as well as precious radio resources for handover[16] measurements and handover signaling even when there is no need to exchange data traffic.

During registration, authentication is carried out. Mutual authentication is quite common in 3G, 4G and 5G. Mutual authentication means that the device authenticates the network and the network authenticates the device. The core network stores important information about the device such as a secret key for authentication. While exact authentication mechanisms depend on the technologies, one popular approach involves an authentication algorithm in the network that uses the subscriber-specific secret key and a random number to generate a response. The core network challenges the device with the random number and the device uses the same authentication algorithm to create a response. The core network then compares the response from the device with the expected response to authenticate the device. To facilitate authentication of the network by the device, the network sends an authentication token and the device compares this token with the expected token to authenticate the network.

[15] Although specific terms used for such procedures are often technology-dependent and there are minor technology-specific variations, these processes are universally carried out in 3G, 4G and 5G cellular technologies.

[16] Handover is a process where the serving cell, sector or even the serving base station for the device is changed due to the mobility of the user to ensure that the device has a dedicated radio connection with the best possible communications link. In addition, handover may be used to balance the load among serving base stations and among carrier frequencies available in a cell or sector.

Once authentication is completed, communications between the device and the network is secured. For example, information can be encrypted so that the original information cannot be retrieved by eavesdropping. If the original information bits are "100111", the transmitter may encrypt these bits and transform them into encrypted bits with an entirely different bit pattern, e.g. "001011". Only the intended recipient can decrypt such encrypted bits and retrieve the original bits. A message authentication code may also be added to the information to ensure that the information has not been tampered with between the transmitter and the receiver. Such a process is called integrity protection. For example, let us say the original set of bits are "100101001". A suitable integrity protection algorithm uses these bits as the input and creates the message authentication code "110" that is a function of these original bits. The transmitter appends "110" to the original information bits and transmits a longer set of bits "100101001110". The receiver uses the received information bits (i.e. "100101001" in the example) to determine the message authentication code. If the code determined by the receiver does not match "110", it means that the original message bits have been altered between the transmitter and the receiver. If the message authentication code determined by the receiver matches the code present in the received payload, then the received payload is valid and has not been altered.

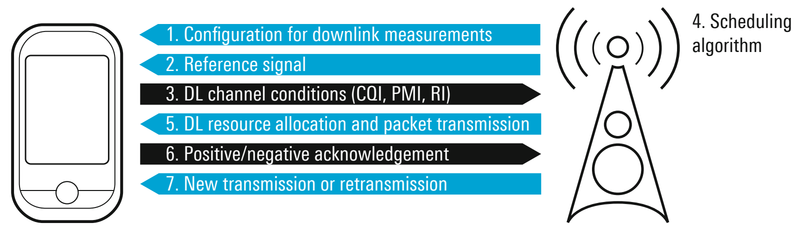
Suitable links are established on the radio interface as well as within the network to facilitate the flow of user traffic. Finally, the network allocates an IP address to the device. In general, an IPv4 or IPv6 address is assigned to the device. The device may be allocated more than one IP address.

All the prerequisites for data transfer have now been met. As soon as the base station allocates radio resources for data transfer, user traffic can start flowing in the downlink and the uplink.

3.5 Data transfer

The radio access network manages the precious radio resources and allocates these resources among to the active devices in a cell. Advanced 3G networks such as 1xEV-DO and HSPA as well as LTE and 5G networks give the base stations the responsibility of managing radio resource. In particular, the MAC layer in the radio protocol stack implements a scheduling algorithm that allocates radio resources to the active devices.
Fig. 23 summarizes the overall steps of downlink data transfer for a typical advanced 3G, 4G or 5G cellular network although the exact details vary from one technology to another.

Fig. 23: Downlink data transfer steps


1. Configuration for downlink measurements
2. Reference signal
3. DL channel conditions (CQI, PMI, RI)
4. Scheduling algorithm
5. DL resource allocation and packet transmission
6. Positive/negative acknowledgement
7. New transmission or retransmission

Prior to any data transfer, the base station configures the device to observe and report the downlink radio channel conditions in **step 1**.

In **step 2**, the base station transmits a cell-specific reference signal or a pilot channel with a known sequence of symbols. The device observes that signal or channel to estimate the quality of radio channel conditions. For example, based on a proprietary implementation-specific approach, the device measures quantities such as the signal-to-interference ratio (SIR) and estimates the supportable data rate at a given target error rate. A target error rate is specified by the standards bodies in the form of a block error rate (BLER) or packet error rate (PER). LTE and HSPA specify an instantaneous target BLER of 10%[17]. 5G allows the error rate to be configurable instead of a fixed value so that QoS requirements can be better met for different services.

In **step 3**, the device reports the estimated downlink channel conditions to the base station in the form of technology-specific quantities such as the channel quality indicator (CQI) for LTE and HSPA and the data rate control (DRC) value for CDMA based 1xEV-DO. Such quantities imply the supportable combination of the modulation and coding scheme (MCS) under given radio channel conditions at the target error rate. For example, if the radio channel conditions are good, a relatively aggressive MCS of 64QAM and coding rate of 2/3 can be used. Since a 64QAM modulation symbol represents 6 code symbols and the coding rate of 2/3 corresponds to two bits becoming 3 code symbols, the effective number of bits represented by a modulation symbol is 6 × 2/3 = 4. Poor channel conditions may require the use of a more robust QPSK modulation scheme and a coding rate of 1/6. The modulation symbol would then represent only 2 × 1/6 = 0.33 bit. These examples show that radio channel conditions have a significant influence on the achievable throughput.

When advanced antenna techniques are supported by the base station and the device (which is a typical scenario for LTE and 5G but not a prevalent scenario for 3G), additional feedback in support of antenna techniques is needed. For example, the device sends feedback such as a precoding matrix indicator (PMI) and rank indicator (RI). PMI specifies how modulation symbols can be transformed through suitable phase shifts prior to the IFFT operation and placed on specific antennas. RI specifies the number of spatial multiplexing layers. If RI = 2, the base station can use 2x2 SU-MIMO for the device to (theoretically) double the throughput compared to single-antenna transmission and single-antenna reception. Similarly, if RI = 4, the base station can use 4x4 SU-MIMO for the device to (theoretically) quadruple the throughput compared to single-antenna transmission and single-antenna reception.

In **step 4**, the base station scheduler processes the reports from active devices on the prevailing downlink radio channel conditions. The scheduler plays an important role in dictating the overall radio access network performance. The scheduler implementation is base station vendor proprietary and a key product differentiator for vendors. The scheduler considers a variety of factors such as the radio channel conditions reported by the active devices (e.g. CQI, PMI and RI), the amounts and types of user traffic in buffers at the base station, target QoS parameters configured for each device and its active services, capabilities of devices and the base station, wait times of packets and available radio resources. The scheduler then selects the users for transmission in the next transmission time interval (TTI) and the transmission parameters for these selected users. Examples of transmission parameters include the amount and type of data to be sent, the antenna technique and associated parameters such as the number of layers, suitable layer mapping and precoding, and radio resources (e.g. physical resource blocks (PRB) in the case of an LTE or 5G network).

In **step 5**, the base station informs the selected devices about allocated downlink radio resources[18] and the devices collect their packets by processing the allocated radio resources.

[17] While the error rate of 10% may seem high, it is an instantaneous error rate and a single retransmission statistically brings the effective error rate down to 1% (i.e. probability of first transmission in error x probability of first retransmission in error = 0.1 × 0.1 = 0.01).

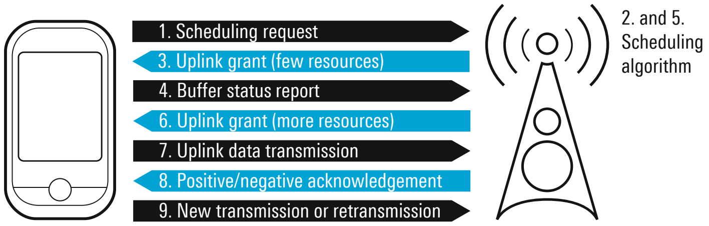
[18] In 1xEV-DO, the base station does not explicitly signal the exact transmission parameters to the device since the base station uses the transmission parameters specified by the device on the uplink DRC channel. In other technologies, including HSPA, LTE and 5G, the base station explicitly conveys transmission parameters to the device.

After the device decodes the packet, in **step 6**, it sends the base station a positive acknowledgment in the case of successful reception and a negative acknowledgment in the case of unsuccessful reception. Such acknowledgements were introduced in the enhanced versions of 3G technologies such as 1xEV-DO and HSPA and later technologies such as LTE and 5G have continued making use of such fast acknowledgments due to their significant positive impact on throughput and subsequently user experience.

In **step 7**, the base station transmits a new packet if the device sent a positive acknowledgment in **step 6** and retransmits the previously sent packet (with possibly a different set of code symbols) if the device sent a negative acknowledgment in **step 6**. The maximum number of such fast retransmissions at the physical layer is typically determined by the base station. After the limit on the number of retransmissions is reached, the physical layer gives up and lets the RLC layer do any retransmissions (if configured). If both the physical layer and the RLC layer cannot correct errors through retransmissions, the transport layer protocol, e.g. transmission control protocol (TCP), does the retransmissions. The physical layer retransmissions are called hybrid automatic repeat request (HARQ) retransmissions and the RLC layer retransmissions are called automatic repeat request (ARQ) retransmissions. The combination of various features such as AMC, HARQ and ARQ enable the wireless network to achieve a very low error rate such as a residual error rate of one packet error out of a million packets (and even one packet error out of a billion packets in 5G).

Now that we have described downlink data transfer, let us turn our attention to uplink data transfer. Uplink and downlink data transfer share several similarities. For example, radio channel conditions are observed, the base station scheduling algorithm allocates radio resources to the device, suitable transmission parameters are used and fast retransmissions occur when needed. Fig. 24 summarizes the overall steps of uplink data transfer for typical LTE and 5G cellular networks. Specific details vary from one technology to another.

Fig. 24: Uplink data transfer steps


1. Scheduling request
3. Uplink grant (few resources)
4. Buffer status report
6. Uplink grant (more resources)
7. Uplink data transmission
8. Positive/negative acknowledgement
9. New transmission or retransmission
2. and 5.
Scheduling
algorithm

When the device has data to send in the uplink, **step 1** is to use a dedicated signaling connection to send a scheduling request to the base station.

In **step 2**, the base station executes a scheduling algorithm and, in **step 3**, allocates limited uplink radio resources to the device so that the device can send a buffer status report in **step 4**. The buffer status report specifies the amount and type of data in the buffer at the device. This report allows the base station to determine the appropriate amount of radio resources to allocate to the device. For example, if the device reports a large amount of data, the base station can allocate more radio resources to the device so that the data can be quickly transferred and the average packet latency can be reduced.

In **step 5**, the base station scheduler processes the buffer status reports from active devices. Like in the downlink, the scheduler plays an important role in dictating the overall radio access network performance. The scheduler considers a variety of factors such as the observed uplink radio channel conditions, the amounts and types of user traffic reported in buffer status reports, target QoS parameters configured for each device and its active services, uplink capabilities of devices and the base station and available radio resources. The scheduler then selects the users for uplink transmission in the next TTI and specifies the uplink transmission parameters (e.g. the modulation and coding scheme combination) and uplink radio resources for these selected users.

In **step 6**, the base station informs the selected devices about allocated uplink radio resources. In **step 7**, the device sends a packet using the allocated uplink radio resources and specified transmission parameters.

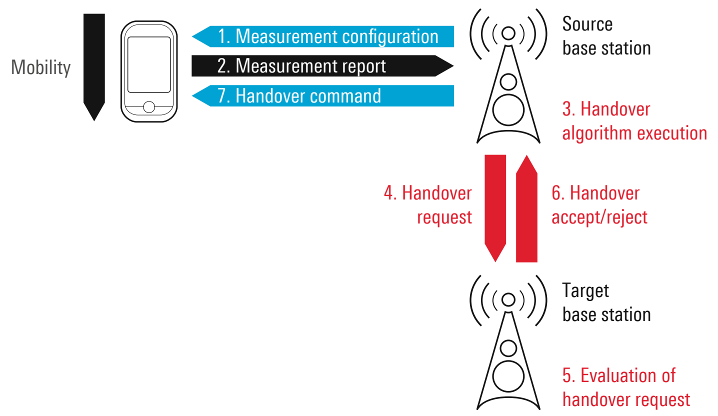
After the base station decodes the packet in **step 8**, it sends the device a positive acknowledgment in the case of successful reception and a negative acknowledgment in the case of unsuccessful reception. The base station also allocates suitable radio resources to the device so that the device can send a new packet in the uplink or retransmit the old packet in **step 9**.

3.6 Handover and cell reselection

As the mobile device moves around in a given area, it crosses cell boundaries. Handover is a process where the dedicated radio connection between the device and the radio access network is switched from one cell to another. In LTE and 5G networks, the device typically communicates with a single cell or sector[19]. In CDMA based networks such as UMTS, 1x and 1xEV-DO, the device simultaneously communicates with multiple cells or sectors. Handover in CDMA based networks implies a change in one or more cells that serve the device. The set of serving cells is called the active set.

Fig. 25 shows the main steps of handover. In **step 1**, the radio access network configures the device for radio channel measurements and measurement reporting. For example, the LTE radio access network may ask the device to make measurements such as reference signal received power (RSRP) and reference signal received quality (RSRQ) on the currently serving cell and neighboring cells. CDMA based networks such as UMTS, HSPA, 1x and 1xEV-DO ask the device to make measurements such as the ratio of the pilot channel power and the total received power[20].

Fig. 25: Overview of handover steps


Mobility
1. Measurement configuration
2. Measurement report
7. Handover command
Source base station
3. Handover algorithm execution
4. Handover request
6. Handover accept/reject
Target base station
5. Evaluation of handover request

[19] In LTE and 5G, a feature called coordinated multi-point (CoMP) allows the device to simultaneously communicate with multiple cells.

[20] The UMTS network calls this E_c/N_0, while the 1x and 1xEV-DO network call it E_c/I_0.

In **step 2**, the device sends a measurement report to the radio access network. Depending upon the technology and the configuration chosen by the radio access network, measurement reporting can be event based and/or periodic. Event based reporting means that the device sends a measurement report when an event occurs. For example, when a neighboring cell measurement is better than the serving cell measurement, an event is said to occur and the device sends a measurement report containing relevant measurements. Periodic reporting, as the name suggests, involves the device sending measurement reports at the interval configured by the radio access network.

In **step 3**, the handover algorithm in the radio access network processes the measurement report received from the device and makes a handover decision if deemed necessary and/or efficient. The radio access network also chooses the target cell for the handover based on the measurement report.

In **step 4**, the currently serving base station contacts the target base station[21] that is managing the target cell to check if the target base station can accommodate the handover request.

In **step 5**, the target base station evaluates the radio resource requirements of the handover request and in **step 6** accepts or rejects the request made by the source base station. The target base station also typically specifies the radio resources that the device can use to contact the target base station upon completion of handover.

In **step 7**, the source base station sends a handover command to the device and the device establishes a radio connection with the target base station. Suitable connections between the target base station and the core network are also set up so that end-to-end data transfer is not interrupted due to handover.

If a device with a dedicated radio connection to the network does not engage in any data transfer for a period of time (e.g. a few seconds or minutes), maintaining the radio connection is quite expensive from the resource consumption perspective. The connected device needs to make frequent measurements on the serving and neighboring cells, and the base station needs to exchange handover related signaling messages with neighboring base stations. To avoid such processing, the network asks the device to enter idle mode after a period of inactivity to save the device's battery power and eliminate unnecessary consumption of precious radio resources. In idle mode, the device does not have a dedicated radio connection.

In idle mode, the device periodically (and less frequently than in active or connected mode) wakes up to monitor the strongest cell, while saving processing power during other times. If the device finds a better cell, it performs cell reselection and starts monitoring the new better cell. The core network knows about the device location at the paging area level, where a paging area contains numerous base stations[22]. In the case of incoming traffic, the core network instructs base stations within the device's paging area to send a page message. The device responds to the page message in the strongest cell that it has been monitoring and exits idle mode to set up a dedicated radio connection. If the idle device wants to send a traffic or signaling message in the uplink, it utilizes the random access channel to contact the base station and exits idle mode. If the device moves from one paging area to another paging area, it carries out a paging area update so that the core network becomes aware of the new paging area for the device.

[21] In 3G networks, the handover algorithm is implemented in the radio network controller (RNC); the RNC manages the resources of multiple base stations.

[22] In LTE and 5G, a paging area is a set of one or more geographic areas called tracking areas.

4 Summary

Cellular networks, independent of generations, share many common attributes end user or handset devices, radio network, core network and services network. Many of the protocol-based communications exchanges between the device and the base station follow a similar philosophy of identifying a potential cell, registering and authenticating with the core network and support for mobility through handover signaling. There are differences in processing mechanisms that occur between different generations of wireless devices, such as how the spectrum is accessed and shared, error correction mechanisms and the way that multipath is handled. These differences were discussed for 3G and 4G systems.

Deployment of 5G systems began in 2018, and we will see many new features for this technology for years to come. Examples of these features include:

- Massive MIMO to increase capacity and throughput and improve link reliability
- A flexible frame structure to support diverse services and industries
- Advanced channel coding schemes such as polar coding and LDPC coding to reduce the error rate and overall power consumption
- Millimeterwave bands (e.g. greater than 24 GHz) with large bandwidth and significant propagation challenges
- Spectrum sharing between cellular technologies and non-cellular technologies (e.g. radar), including sharing with devices using unlicensed spectrum
- Low latency, low power, highly reliable wireless connections in support of IoT applications
- Ability to support much greater density in the deployment of user devices, especially to support IoT
- Multi-access edge computing (MEC) to reduce end-to-end latency and bandwidth requirements for the transport network
- Network slicing to realize custom and comprehensive QoS for diverse services and customer requirements
- Virtualization and automation technologies such as network function virtualization (NFV) and software defined networking (SDN) to implement cloud based radio access networks, core networks and service networks for service agility, enhanced service experience, scalability and cost-effective, end-to-end network design, maintenance and optimization processes

Rapidly changing wireless technologies are challenging for the industry and bring a new level of respect for the challenges faced by equipment manufacturers. Test equipment has to be available before mass deployment of new wireless technologies, which means that equipment manufacturers must be two steps ahead of the wireless industry in understanding and implementing new technologies. Designing test equipment that assesses the performance of systems based on new concepts is not the only goal. The test equipment is only as good as the understanding of the engineer using the equipment. At Rohde & Schwarz, our mission is not only to sell leading-edge test equipment but also to help provide the knowledge and training that our customers need to handle the ever-increasing technical challenges of emerging wireless systems.

5 References

[1] Nishith D. Tripathi and Jeffrey H. Reed, "Cellular Communications: A Comprehensive and Practical Guide", IEEE/Wiley, 2014

[2] 3GPP, TS 36.101, "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception"

[3] 3GPP, TS 38.101, "NR; User Equipment (UE) radio transmission and reception"

[4] Nishith D. Tripathi and Jeffrey H. Reed, "5G Cellular Communications: Journey and Destination", multimedia eBook, to be published in fall 2018.

[5] Stephen Temple, Inside the Mobile Revolution – A Political History, 2nd Edition, 2010. http://www.gsmhistory.com/wp-content/uploads/2013/01/Inside-a-Mobile-Revolution-Temple-20101.pdf

Rohde & Schwarz

The Rohde & Schwarz electronics group offers innovative solutions in the following business fields: test and measurement, broadcast and media, secure communications, cybersecurity, monitoring and network testing. Founded more than 80 years ago, the independent company which is headquartered in Munich, Germany, has an extensive sales and service network with locations in more than 70 countries.

www.rohde-schwarz.com

Regional contact

- Europe, Africa, Middle East | +49 89 4129 12345
  customersupport@rohde-schwarz.com
- North America | 1 888 TEST RSA (1 888 837 87 72)
  customer.support@rsa.rohde-schwarz.com
- Latin America | +1 410 910 79 88
  customersupport.la@rohde-schwarz.com
- Asia Pacific | +65 65 13 04 88
  customersupport.asia@rohde-schwarz.com
- China | +86 800 810 82 28 | +86 400 650 58 96
  customersupport.china@rohde-schwarz.com

PD 5216.0467.52 | Version 02.01 | January 2019 (sk)
White paper | Radio fundamentals for cellular networks
Data without tolerance limits is not binding | Subject to change

5216046752

5216.0467.52 02.01 PDP 1 en