UNIT-I (INTRODUCTION TO 5G)

3G and 4G(LTE) overview- Introduction to 5G – Use Cases - Evolving LTE to 5G Capability- 5G NR and 5G core network (5GCN) - 5G Standardization - 3GPP and IMT2020 - Spectrum for 5G – 5G deployment - Options, Challenges and Applications.

1. Introduction :

Mobile Cellular Telephony is one of the greatest innovations of the twentieth century and today in the twenty-first century it can be safely said that it has brought nothing less than a revolution in the way communications take place across the globe. The Mobile Cellular Telephony is enabled through a combination of cellular networks and mobile devices which communicate to each other by means of radio frequency spectrum (i.e. wirelessly). A cellular network consists of thousands of nodes that assist mobile device users in performing plethora of tasks. Mobile device has become the Third Screen after Television and Computer and is becoming more economical and powerful with continual technological advancements. There are more than 5 billion mobile subscribers in the world and for the very vast majority the mobile device has become a necessity and without it they can't go about with their daily routine lives.

1.1 Mobile Cellular Telephony Evolution

The cellular concept was conceived by Bell Laboratories in 1947 and enabled companies to provide wireless communications to a large population [1]. Like any other field of science and technology, mobile communications is continuously evolving and the sector* has made astonishing progress in the last 70 years. The first generation (1G) cellular networks was deployed in the 1980s, the second generation (2G) in the 1990s, while the third generation (3G) in the 2000s. Today, 4G (fourth generation) cellular networks are being deployed and the world is getting ready to embrace the fifth generation (5G) of mobile cellular telephony. The 1G analog system are no longer operational, which only provided voice services and had no support for data. The 2G digital systems are currently operational and support voice and limited data services. The 3G systems support voice, low speed data, and enable a number of data services. The 4G systems enable mobile broadband in the true sense, targeting 100 Mbps or higher on the move. 5G systems are expected to provide an enhanced mobile broadband targeting peak data rate of 20 Gbps, extend 4G's Internet of Things capability, and enable mission-critical applications that require ultra-high reliability and low latency. 5G networks are expected to be designed by taking a user-centric approach.

Simply, the "G" stands for "GENERATION". While connected to the internet, the speed of the connection depends upon the signal strength that is shown in abbreviations like 2G, 3G, 4G, 5G, etc. on any mobile device. Each generation of wireless broadband is defined as a set of telephone network standards that describe the technological implementation of the system. The aim of wireless communication is to provide high quality, reliable communication just like wired communication and each new generation represents a big leap in that direction. Mobile

communication has become more popular in the last few years due to fast reform in mobile technology. For the comparison of 2G, 3G, 4G, and 5G we first need to understand the key features of all these technologies.

Second Generation (2G): 2G refers to the second generation of mobile networks based on GSM. The radio signals used by the 1G network were analog, while 2G networks were digital. 2G capabilities were achieved by allowing multiple users on a single channel via multiplexing. During 2G, cellular phones were used for data along with voice. Some of the key features of 2G were: Data speeds of up to 64 kbps Use of digital signals instead of analog Enabled services such as SMS and MMS (Multimedia Message) Provided better quality voice calls It used a bandwidth of 30 to 200 KHz.

Third Generation (3G): The 3G standard utilises Universal Mobile Telecommunications System (UMTS) as its core network architecture. 3G network combines aspects of the 2G network with new technologies and protocols to deliver a significantly faster data rate. By using packet switching, the original technology was improved to allow speeds up to 14 Mbps. It used Wide Band Wireless Network that increased clarity. It operates at a range of 2100 MHz and has a bandwidth of 15-20 MHz. Some of the main features of 3G are: Speed of up to 2 Mbps Increased bandwidth and data transfer rates Send/receive large email messages Large capacities and broadband capabilities International Mobile Telecommunications-2000 (IMT-2000) were the specifications by the International Telecommunication Union for the 3G network; theoretically, 21.6 Mbps is the max speed of HSPA+.

Fourth Generation (4G): The main difference between 3G and 4G is the data rate. There is also a huge difference between 3G and 4G technology. The key technologies that have made 4G possible are MIMO (Multiple Input Multiple Output) and OFDM (Orthogonal Frequency Division Multiplexing). The most important 4G standards are WiMAX and LTE. While 4G LTE is a major improvement over 3G speeds, it is technically not 4G. What is the difference between 4G and LTE? Even after it was widely available, many networks were not up to the required speed of 4G. 4G LTE is a “fourth generation long term evolution”, capable of delivering a very fast and secure internet connection. Basically, 4G is the predetermined standard for mobile network connections. 4G LTE is the term given to the path which has to be followed to achieve those predefined standards. Some of the features of 4G LTE are: Support interactive multimedia, voice, video. High speed, high capacity and low cost per bit (Speeds of up to 20 Mbps or more.) Global and scalable mobile networks. Ad hoc and multi-hop networks.

Fifth Generation (5G): 5G networks operate on rarely used radio millimeter bands in the 30 GHz to 300 GHz range. Testing of 5G range in mmWave has produced results approximately 500 meters from the tower. Using small cells, the deployment of 5G with millimetre wave based carriers can improve overall coverage area. Combined with beamforming, small cells can deliver extremely fast coverage with low latency. Low latency is one of 5G’s most important features. 5G uses a scalable orthogonal frequency-division multiplexing (OFDM) framework. 5G benefits greatly from this and can have latency as low as one millisecond with realistic estimates to be around 1 – 10 seconds. 5G is estimated to be 60 to 120 times faster than the average 4G latency. Active antenna 5G encapsulated with 5G massive MIMO is used for providing better

connections and enhanced user experience. Big 5G array antennas are deployed to gain additional beamforming information and knock out propagation challenges that are experienced at mmWave frequency ranges. Further, 5G networks clubbed with network slicing architecture enables telecom operators to offer on-demand tailored connectivity to their users that is adhered to Service Level Agreement (SLA). Such customised network capabilities comprise latency, data speed, latency, reliability, quality, services, and security. With speeds of up to 10 Gbps, 5G is set to be as much as 10 times faster than 4G.

Comparison	2G	3G	4G	5G
Introduced in year	1993	2001	2009	2018
Technology	GSM	WCDMA	LTE, WiMAX	MIMO, mm Waves
Access system	TDMA, CDMA	CDMA	CDMA	OFDM, BDMA
Switching type	Circuit switching for voice and packet switching for data	Packet switching except for air interference	Packet switching	Packet switching
Internet service	Narrowband	Broadband	Ultra broadband	Wireless World Wide Web
Bandwidth	25 MHz	25 MHz	100 MHz	30 GHz to 300 GHz
Advantage	Multimedia features (SMS, MMS), internet access and SIM introduced	High security, international roaming	Speed, high speed handoffs, global mobility	Extremely high speeds, low latency
Applications	Voice calls, short messages	Video conferencing, mobile TV, GPS	High speed applications, mobile TV, wearable devices	High resolution video streaming, remote control of vehicles, robots, and medical procedures

Figure 1.1 Comparison of 2G, 3G, 4G, and 5G

1.2 Hexagon Based Mobile Cellular Telephony

A cellular network or mobile network is a wireless network spread over the land through a web of cell sites. Each of these sites or cell towers is comprised of a transceiver (transmitter/receiver) for communications with mobile devices. From a technological perspective, mobile devices† rely on die hard cellular towers for communications and these cell sites or cell towers are designed to keep a hexagonal shape in mind. The use of hexagonal cells was invented by Bell Laboratories in the 1970s . This shape was selected over other geometrical shapes since by using it the cells can be laid next to each other with no overlap, thus providing coverage theoretically to the entire service area without any gaps. The hexagon design has been at least so far remained as necessary for mobile communications as cement for the construction of buildings or coal tar for carpeting the roads.

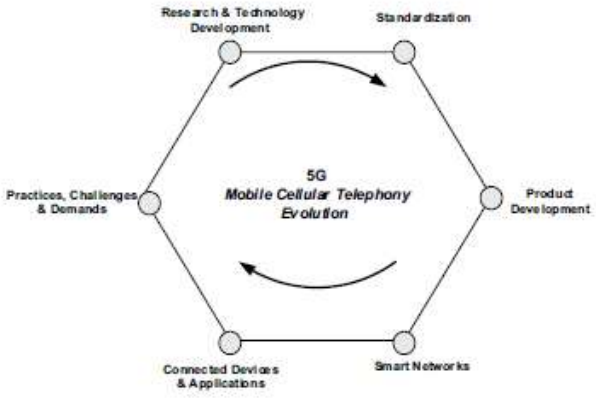


Figure 1.2: Key steps in Mobile Cellular Telephony

If we look at it from an end-to-end perspective, the key phases of the mobile cellular telephony can nicely fit on the six corners of the hexagon shape (Figure 1.1). These generic six phases are research and technology development, standardization, product development, network development, device and application development, and the sector's practices and challenges. In a nutshell, though not necessary, research and technological development is the first step that leads into standardization followed by product development. Once telecom products are ready, they get deployed in the cellular networks. Device development usually lags behind network equipment production. Once networks are up and running and users are connected to networks through their devices, applications start to pour in; from there, the sectors begin to see the good or not-so-good practices, bottle necks and challenges, and new business demands which then leads back to the first step to start all over again.

1.3 Radio Spectrum And Mobile Communications

Today's 2G/3G/4G mobile communications primarily use frequencies in the range of 700 MHz to 42 GHz. Additionally, some communication take place in the 400 MHz and 70/80 GHz range, however, the use of this set of frequencies is relatively very small (just like a needle in a haystack). The frequencies are allocated by the International Telecommunications Union Radio Communication Sector (ITU-R) through World Radio Communication Conferences (WRC) on both a primary and secondary basis. On a wider scale, spectrum sharing of a primary allocation with other primary and/or secondary services has not been attempted as such. For example, the 698–806 MHz band had been historically allocated by ITU-R on a primary basis for both broadcasting and mobile use, but it was only used for broadcasting and not for mobile use in the U.S. (i.e., no sharing was taking place). During 2008–09, the Federal Communications Commission or FCC auctioned this band for mobile communications while ceasing television broadcasting in the same range, thus reducing the opportunity for any sharing between the two services.

5G envisions the use of high capacity broadband applications and services that will require a huge amount of spectrum. Beyond excessive mobile broadband and gigabits of data rates, applications like the Internet of Things, use of wireless sensors, and so on have necessitated the search for additional spectrum. It is widely established that the world will need an additional 1000 or so MHz to meet the demands of mobile broadband by 2020. Spectrum scarcity is emerging as one of key problems for 5G and so far the world has not found a solution or solutions (as readers will see later in the chapter). The ITU through the World Radio Communication Conference 2015 (WRC-15) only allocated 51 MHz for IMT (International Mobile Telecommunications) on a global scale, which is quite infinitesimal compared to what is required. However, the conference has identified several bands in the range of 24.25–86 GHz for studies to address this requirement. The sharing and compatibility studies of these bands are expected to be shared during WRC-19; thus we can expect to find additional spectrum for 5G and broadband in due time.

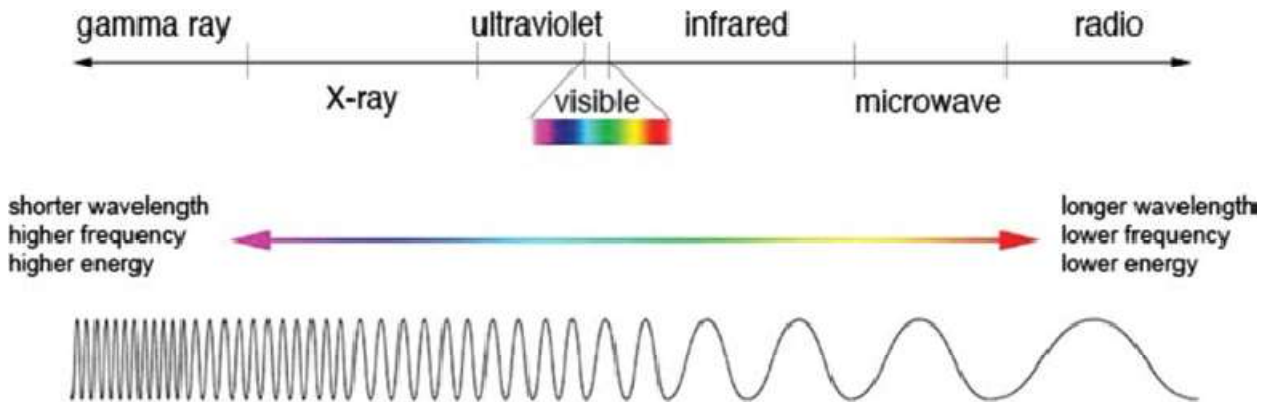


Figure 1.3: Electromagnetic spectrum. (From NASA)

Furthermore, innovative techniques are needed to introduce spectrum sharing and effectively manage implications of air-interface design. And spectrum also has to be managed and regulated to avoid interference related bottlenecks.

1.3.1 Frequency Allocation and Identification

The ITU-R is the body that identifies frequency bands for almost any type of wireless communications all around the world. These types include but are not limited to aviation, broadcasting, maritime, mobile communications, public protection and disaster relief, satellite services, and so on. The allocation and identification of frequencies take place at the ITU WRCs. These inter-governmental events take place every three to four years to address the frequency related needs of the world. WRC is the most significant conference related to the frequency spectrum organized by ITU with a mandate to review, and, if necessary, to revise the radio regulations which govern the use of a scarce resource, that is the frequency spectrum. As far as mobile communications are concerned, the frequencies have to be allocated primarily for mobile communications and may have to be identified for International Mobile Telecommunications (IMT). Second, these have to be allocated for fixed wireless communications links to support backhaul in mobile networks which is termed as a fixed service in the Table of Frequency Allocations (TFA) maintained by ITU at the international level. In simple terms, frequencies have to be identified for wireless communications that take place between mobile users and cellular towers and between cellular towers (i.e., terrestrial backhaul using microwave radios). Finally, satellite based links are also used to support backhaul traffic in remote and far flung areas. Thus, some frequency bands allocated for satellite communications are used in mobile networks in the form of VSAT (Very Small Aperture Terminal), but in the whole scheme of things, their relative use is quite small compared to the frequencies that are used for communications between users and cell phone towers and in terrestrial backhaul.

1.3.2 Frequency Spectrum Needs of 5G

It is well known that the availability of new spectrum bands is a key requirement for the provision of 5G or IMT-2020 services. The ITU-R has estimated that the total global spectrum requirements for IMT will be in the range of 1340 (for lower user density settings) to 1960 MHz (for higher user density settings) for the year 2020 [8]. Spectrum allocation is required not only

in air-interface, but also for backhaul and to some extent in the fronthaul. Fronthaul is the link between a pool of base band units and remote radio units (RRUs) which collectively formed the concept of C-RAN (cloud/centralized radio access network). Backhaul (first leg between RRUs and Core Network) is a major challenge for 5G, but to some extent, it can be fulfilled with wired media such as optical fiber cable and technologies such as very-high-bit-rate digital subscriber line 2, and so on. However, for the most part, the air-interface (link between wireless user/device and remote radio unit) is where the vast majority of the spectrum is required.

1.3.3 Spectrum Sharing

Spectrum sharing is defined as the collective use of a frequency band by two or more parties in a specific geographical area. Sharing can take place in both licensed and license-exempt bands [7]. For 5G, sharing may also need to be considered with incumbents such as FSS (fixed satellite service), radar, and so on. This form of sharing requires compatibility studies between broadband and nonmobile incumbents which ITU-R is planning to complete in conjunction with its preparation for WRC-19. Spectrum sharing is a three-dimensional challenge that not only needs to allow for frequency but must also encompass time and geographical factors in providing access across multiple classes of users. The Licensed Shared Access (LSA) mechanism allows LSA licensees to access the spectrum that has already been assigned to an incumbent. This method allows sharing based on certain rules guaranteeing some level of QoS (Quality of Service). It is different from certain cognitive approaches that allow access to TV white space on an unlicensed basis without any QoS guarantees. LSA allows mobile operators to use/obtain additional spectrum on a secondary basis and with guaranteed access for an agreed geographic area, time frame, and frequency range. Furthermore, licensed sharing can be horizontal, which normally involves sharing between two similar parties (like two mobile operators) whereas with vertical sharing, the frequency can be shared between different types of parties (like mobile operators and a government organization).

All of these licensed sharing options can take place within a specified area (geographic sharing), at specific or random times (temporal sharing), and these need to be coordinated as well to avoid harmful interference. The license-exempt approach, as the name suggests, allows sharing among parties without requiring a license. It enables best effort access and operations for data offloading and so on, and thus it is not well suited for carrier grade performance. The value of a spectrum depends on the profitability of the services that have been assigned to by the respective spectrum. For example, the spectrum assigned for mobile telecommunications will yield much higher economic value than the one that has been assigned to a government entity. Sharing involves tradeoffs whereby allowing a new user in the band will likely diminish what an existing subscriber can do and capitalize on. In general terms, spectrum sharing creates costs and restricts revenues compared to the exclusive use of the same frequency band. Many case studies are available that have determined the impact of sharing on the value of a spectrum. The studies presented in have shown that sharing reduces the economic value of the spectrum.

1.3.4 Air-Interface Design

A primary question, that is, whether there will be a single air-interface or a collection of air-interfaces for 5G, still remains to be resolved. However, for frequency agility, sharing, coexistence, and scalable spectral efficiency, an effective air-interface design is a fundamental enabler and there are several reasons for this. First, there are expected to be several bands for 5G distributed over a large range of frequencies so the air-interface has to be flexible enough to accommodate all such bands. These bands could be contiguous/non-contiguous and can fall anywhere from sub 1 GHz up to 100 GHz. Second, the air-interface has to deal with all the possible spectrum sharing scenarios. The sharing can take place in both licensed and unlicensed bands, with or without the involvement of 5G networks. The main challenge that it needs to manage concerning sharing is interference while also optimizing the efficiency of the spectrum. The interference can be managed in various dimensions including time intervals, orthogonal/nonorthogonal frequency resources, locations with sufficient separation, spatial, and orthogonal codes. Last but not least are the significant variations in uplink and downlink traffic ratios (that even exist today), which imply the need for a flexible air-interface design to effectively manage traffic asymmetry.

1.3.5 Spectral Efficiency

The technical measure of efficiency for a frequency spectrum is called spectral efficiency, which is measured in bits/sec/Hz. Spectral efficiency refers to the information rate that can be transmitted over a given bandwidth in a specific communication system. Numerous results are present on the Internet showing the efficiency of various mobile technologies. Simulation versus on-ground results sometimes differ to some extent as most simulations are conducted in close to ideal conditions, whereas there are practical considerations in implementing technologies in the field [9,18]. Peak and average spectral efficiencies differ due to changing radio conditions. Peak spectral efficiency is calculated using the highest throughput per sector achieved with the combination of a high order modulation scheme, low code rate, and at a high SNR (signal-to-noise ratio) in a given amount of spectrum.

In practical terms, average spectral efficiency is the better unit of measurement which considers aggregate cell throughput per sector within the assigned spectrum. Beside modulation and coding, there are other factors that can impact the efficiency of the spectrum. These include mobile receive diversity, MIMO antenna technology, equalization, and so on. Furthermore, the spectral efficiency can be improved by using wider radio channels. LTE provides roughly 5% better spectral efficiency with a 20 MHz channel as compared to a 10 MHz channel. It may be noted that the spectral efficiency of a radio technology is independent of the frequency it uses to operate, since modulation, coding, and antenna diversity remain the same at different frequencies.

1.4 Evolving LTE to 5G Capability

5G radio access technology will be a key component of the Networked Society. It will address high traffic growth and increasing demand for high-bandwidth connectivity. It will also support massive numbers of connected devices and meet the real-time, high-reliability

communication needs of mission-critical applications. 5G will provide wireless connectivity for a wide range of new applications and use cases, including wearables, smart homes, traffic safety/control, critical infrastructure, industry processes and very-high-speed media delivery. As a result, it will also accelerate the development of the Internet of Things. ITU Members including key industry players, industry forums, national and regional standards development organizations, regulators, network operators, equipment manufacturers as well as academia and research institutions together with Member States, gathered as the working group responsible for IMT systems, and completed a cycle of studies on the key performance requirements of 5G technologies for IMT-2020.

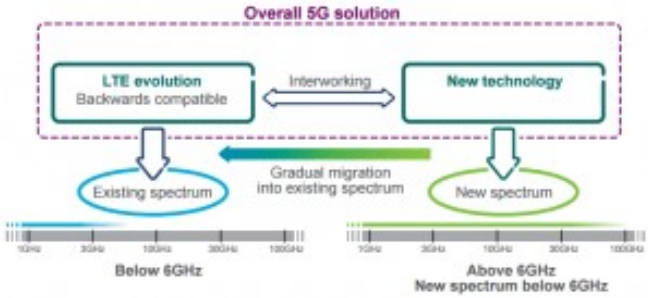


Figure 1.4: Cable free 5G Technology

The specification of 5G will include the development of a new flexible air interface, NX, which will be directed to extreme mobile broadband deployments. NX will also target high-bandwidth and high-traffic-usage scenarios, as well as new scenarios that involve mission-critical and realtime communications with extreme requirements in terms of latency and reliability. In parallel, the development of Narrow-Band IoT (NB-IoT) in 3GPP is expected to support massive machine connectivity in wide area applications. NB-IoT will most likely be deployed in bands below 2GHz and will provide high capacity and deep coverage for enormous numbers of connected devices. Ensuring interoperability with past generations of mobile communications has been a key principle of the ICT industry since the development of GSM and later wireless technologies within the 3GPP family of standards.

In a similar manner, LTE will evolve in a way that recognizes its role in providing excellent coverage for mobile users, and 5G networks will incorporate LTE access (based on Orthogonal Frequency Division Multiplexing (OFDM)) along with new air interfaces in a transparent manner toward both the service layer and users. Around 2020, much of the available wireless coverage will continue to be provided by LTE, and it is important that operators with deployed 4G networks have the opportunity to transition some – or all – of their spectrum to newer wireless access technologies. For operators with limited spectrum resources, the possibility of introducing 5G capabilities in an interoperable way – thereby allowing legacy devices to continue to be served on a compatible carrier – is highly beneficial and, in some cases, even vital. At the same time, the evolution of LTE to a point where it is a full member of the 5G family of air interfaces is essential, especially since initial deployment of new air interfaces may not operate in the same bands.

The 5G network will enable dual-connectivity between LTE operating within bands below 6GHz and the NX air interface in bands within the range 6GHz to100GHz. NX should also allow for user-plane aggregation, i.e. joint delivery of data via LTE and NX component carriers. This paper explains the key requirements and capabilities of 5G, along with its technology components and spectrum needs. In order to enable connectivity for a very wide range of applications with new characteristics and requirements, the capabilities of 5G wireless access must extend far beyond those of previous generations of mobile communication. These capabilities will include massive system capacity, very high data rates everywhere, very low latency, ultra-high reliability and availability, very low device cost and energy consumption, and energy-efficient networks.

1.4.1 Massive System Capacity:

Traffic demands for mobile-communication systems are predicted to increase dramatically. To support this traffic in an affordable way, 5G networks must deliver data with much lower cost per bit compared with the networks of today. Furthermore, the increase in data consumption will result in an increased energy footprint from networks. 5G must therefore consume significantly lower energy per delivered bit than current cellular networks. The exponential increase in connected devices, such as the deployment of billions of wirelessly connected sensors, actuators and similar devices for massive machine connectivity, will place demands on the network to support new paradigms in device and connectivity management that do not compromise security. Each device will generate or consume very small amounts of data, to the extent that they will individually, or even jointly, have limited impact on the overall traffic volume. However, the sheer number of connected devices seriously challenges the ability of the network to provision signaling and manage connections.

1.4.2 Very high Data Rates Everywhere:

Every generation of mobile communication has been associated with higher data rates compared with the previous generation. In the past, much of the focus has been on the peak data rate that can be supported by a wireless-access technology under ideal conditions. However, a more important capability is the data rate that can actually be provided under real-life conditions in different scenarios.

- 5G should support data rates exceeding 10Gbps in specific scenarios such as indoor and dense outdoor environments.
- Data rates of several 100Mbps should generally be achievable in urban and suburban environments.
- Data rates of at least 10Mbps should be accessible almost everywhere, including sparsely populated rural areas in both developed and developing countries.

1.4.3 Very Low Latency:

Very low latency will be driven by the need to support new applications. Some envisioned 5G use cases, such as traffic safety and control of critical infrastructure and industry processes, may require much lower latency compared with what is possible with the mobile-communication systems of today. To support such latency-critical applications, 5G should allow

for an application end-to-end latency of 1ms or less, although application-level framing requirements and codec limitations for media may lead to higher latencies in practice. Many services will distribute computational capacity and storage close to the air interface. This will create new capabilities for real-time communication and will allow ultra-high service reliability in a variety of scenarios, ranging from entertainment to industrial process control.

1.4.4 Ultra-High Reliability And Availability:

In addition to very low latency, 5G should also enable connectivity with ultra-high reliability and ultra-high availability. For critical services, such as control of critical infrastructure and traffic safety, connectivity with certain characteristics, such as a specific maximum latency, should not merely be 'typically available.' Rather, loss of connectivity and deviation from quality of service requirements must be extremely rare. For example, some industrial applications might need to guarantee successful packet delivery within 1 ms with a probability higher than 99.9999 percent.

1.4.5 Very Low Device Cost And Energy Consumption:

Low-cost, low-energy mobile devices have been a key market requirement since the early days of mobile communication. However, to enable the vision of billions of wirelessly connected sensors, actuators and similar devices, a further step has to be taken in terms of device cost and energy consumption. It should be possible for 5G devices to be available at very low cost and with a battery life of several years without recharging.

1.4.6 Energy-Efficient Networks

While device energy consumption has always been prioritized, energy efficiency on the network side has recently emerged as an additional KPI, for three main reasons:

- Energy efficiency is an important component in reducing operational cost, as well as a driver for better dimensioned nodes, leading to lower total cost of ownership.
- Energy efficiency enables off-grid network deployments that rely on medium-sized solar panels as power supplies, thereby enabling wireless connectivity to reach even the most remote areas.
- Energy efficiency is essential to realizing operators' ambition of providing wireless access in a sustainable and more resource-efficient way.

The importance of these factors will increase further in the 5G era, and energy efficiency will therefore be an important requirement in the design of 5G wireless access.

1.5 5G NR(New Radio):

The 5G NR is currently under the defining stage. According to ITU-R, IMT-2020 (including 5G) are mobile systems that include new radio interface(s) that support new capabilities of systems beyond IMT-2000 and IMT-Advanced . The ITU-R has envisioned the following three usage scenarios for IMT-2020

• eMBB: providing higher speeds for applications such as web browsing, streaming, and video conferencing.

• URLLC: enables mission-critical applications, industrial automation, new medical applications, and autonomous driving that require very short network traversal times.

• mMTC: extends LTE IoT capabilities to support a huge number of devices with enhanced coverage and long battery life.

1.5.1 Requirements:

The ITU-R has published a schedule for the standardization of IMT-2020 and set the following timelines

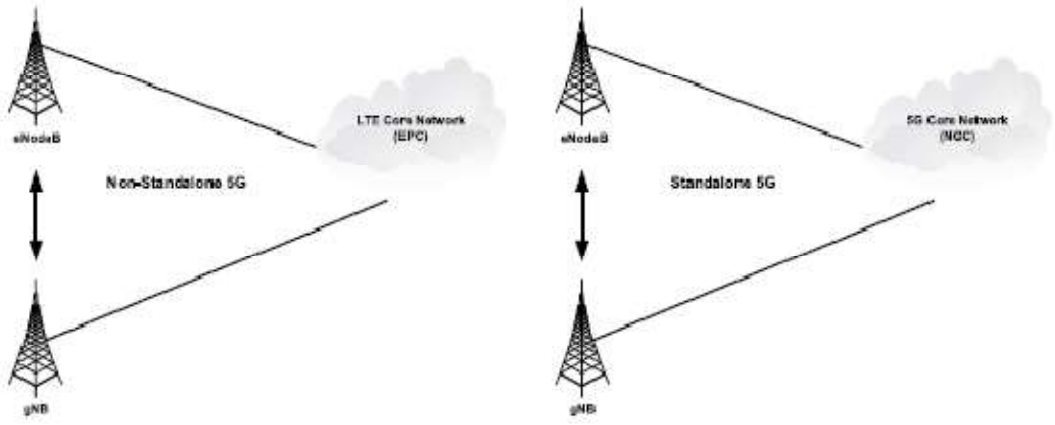
• 2016–2017: Determine technical performance requirements, evaluation criteria, and identification of spectrum bands.

• 2018–2019: Submission and evaluation of proposal.

• 2019–2020: Release of IMT-2020 based radio specifications.

In line with its schedule, the ITU-R published/approved minimum technical performance criteria for IMT-2020 candidate radio interface technologies in 2017. It defines eight key "Capabilities for IMT-2020," which forms a basis for the 13 technical performance requirements were presented. The same recommendation also acknowledges that these key capabilities will have different relevance and applicability for the various usage scenarios of IMT-2020. In comparison to IMT-Advanced (4G), the requirements are tougher and more diversified. For instance, downlink peak data is 1 Gbps in 4G while it is 20 Gbps in IMT-2020, recommended control plane latency is 100 ms in 4G while it is 20 ms in IMT-2020, and many other changes.

3GPP is expected to be the key SDO to answer ITU-R's call for proposals on IMT-2020. 3GPP is a well-known consortium of seven SDOs, namely, ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, and TTC, known as Organizational Partners and a few industry forums. 3GPP has over 500 members (commercial organizations as well as academia) associated with one of these organizational partners. So, in a nutshell, all the key stakeholders of ICT as well representatives from the transportation, energy and other sectors are present in 3GPP, working to improve/innovate the telecom landscape.



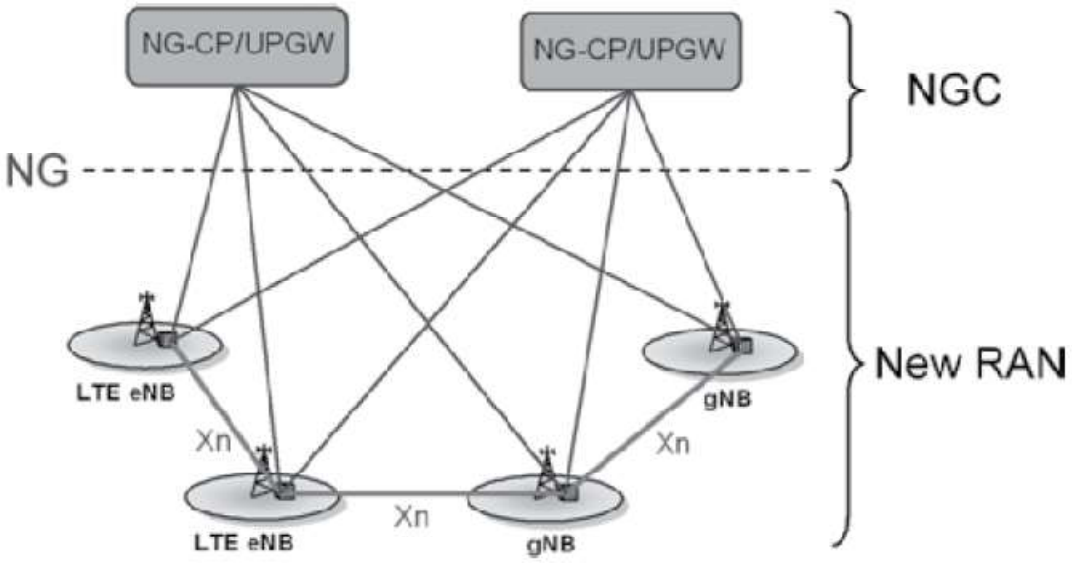


Figure 1.5 : Overall layer 2 structure for NR

1.5.2 5G RAN Architecture:

5G RAN will consist of logical nodes, namely, gNBs and eNBs, that are connected to each other with an Xn interface and toward the NGC by means of an NG interface. The Xn interface connects two gNBs or two eNBs or one gNB with an eNB. The Xn control plane interface (Xn-C) and Xn user plane interface (Xn-U) are defined between two NEW RAN nodes for respective purposes. The NG interface supports signalling between gNB/eNB and NGC, control and user planes' separation, and many other functions. This interface is divided into control and user planes' interfaces, namely, NG control plane interface (NG-C) and NG user plane interface (NG-U). The former is defined between NR gNB/eNB and NG-CPGW (NG control plane gateway) while the latter is identified between gNB/eNB and NG-UPGW (NG user plane gateway). The NG interface supports one-to-many relationships between NGC and new RAN nodes.

The protocol stack of these interfaces is shown in Figure 1.6. The SCTP (Stream Control Transmission Protocol) layer sits on top of the IP/Transport layer, providing guaranteed delivery of application layer messages in the control plane. In the user plane, GTP-U (GPRS tunneling protocol user plane) provides nonguaranteed delivery of PDUs between the respective network elements. Lastly, the application layer signalling protocol is defined as Xn-AP (Xn Application Protocol).

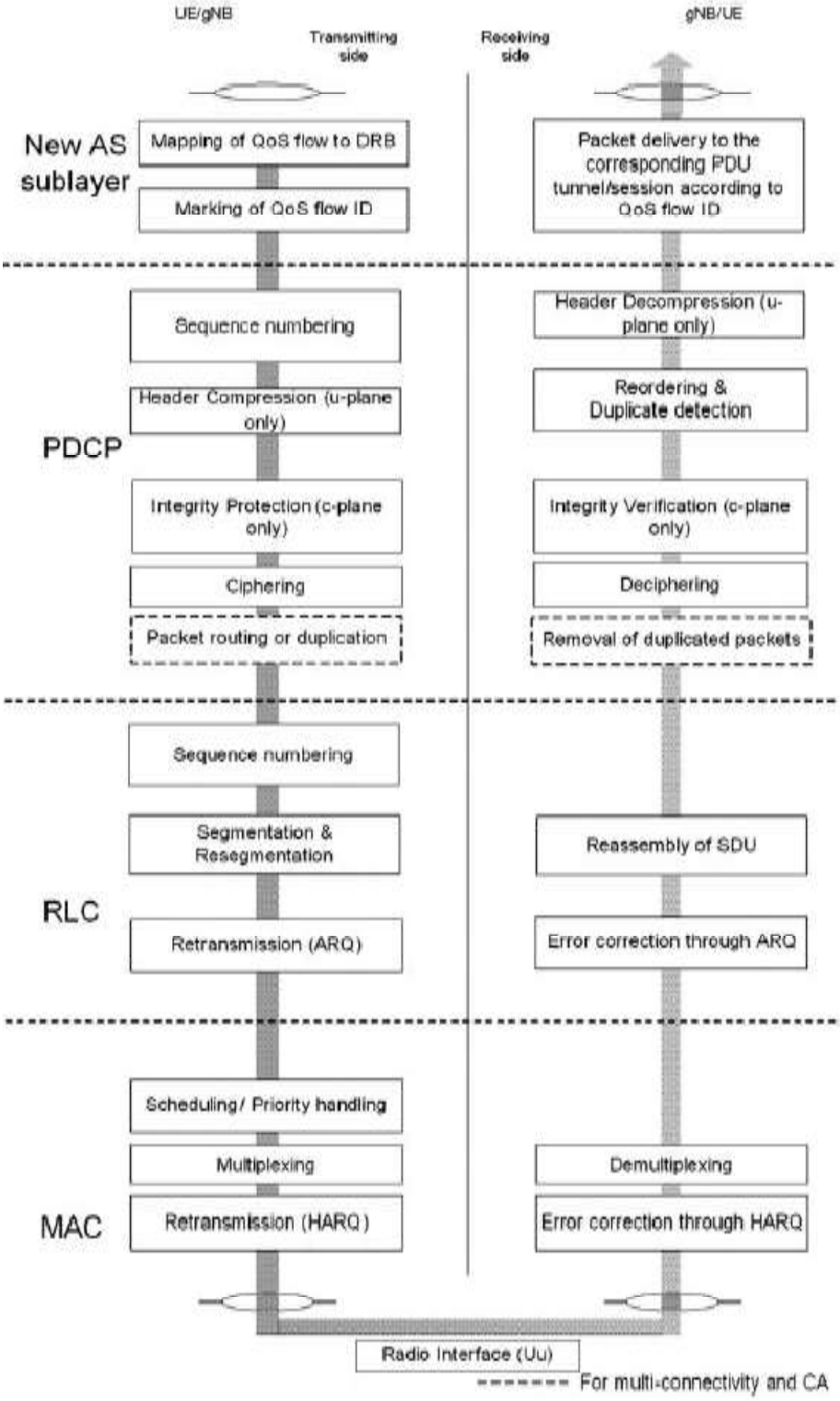


Figure 1.6: Overall layer 2 structure

1.6 5G Core Network:

Core network is the segment that connects the RAN with operators' in-house application servers, IP multimedia subsystem or the Internet. It is comprised of both circuit switched and packet switched elements. Historically, it supported only circuit switched (CS) services, but later with the advent of 3G it started to support packet switched (PS) services as well. LTE and 4G systems only require PS support and with the passage of time, core networks are expected to only provide for IP/Ethernet services. Figure 11.1 describes the evolution of core networks from 2G to 5G based on 3GPP standards. The initial 2G systems supported CS only with key elements like MSC and SMSC, and common elements HLR, VLR, EIR, and SGW*. To support data services, SGSN and GGSN were introduced in 3GPP Release-99 which was the first release on UMTS (3G). Release-4 (2001) divided the MSC into two functional elements, namely, an MSC server for signalling and a media gateway function as the user plane to reduce operational challenges. Later, Release-5 (2002) brought IMS, which was primarily developed for the mobile 3G devices communicating over IP with embedded SIP clients. In Release-6 (2005), a new functional node, that is, BM-SC (the Broadcast Multicast-Service Center), was added to support MBMS. Release-7 (2007) introduced the concept of direct tunneling which enables a split between the control plane and user plane toward the packet core networks.

LTE (Release 8) brought EPC that was only designed to support PS services including elements like MME to manage user equipment mobility and identity and Gateway (Serving and Packet) for packet routing and connecting to external networks, respectively. Rel-9, Rel-10 (LTE-Advanced/4G), Rel-11, and Rel-12 have not brought any fundamental architectural changes in EPC. Rel-13 has introduced the concept of Dedicated Core Network (DCN) along with network slicing which will be explained later in the chapter. The specifications on the 5G core network are expected to be finalized in Rel-15 (2018) and Rel-16 (2020). For 5G, we may see EPC or core network going in the Cloud supported by technologies such as SDN and NFV, addressing all types of IP based services. The chapter provides a short overview of EPC, the evolution of EPC, and the evolution of IMS, whereas the details on core networks can be found in [1]. Additionally, this chapter also briefly presents insights on 5G core network, CDN (Content Delivery Network), and LTE and 5G OSS/BSS (Operational/Business Support Systems).

1.6.1 Evolved Packet Core:

3GPP Rel-8 produced EPC to support LTE and only packet switched traffic. The EPC consists of two main entities, namely MME and Gateway as illustrated. MME supports the control plane functionalities like SSGN, and it is separated from the node that performs bearer-plane functionality, that is, Gateway (GW). Key functions of EPC are as follows: The MME main function is to manage UE mobility and UE identity. It also performs authentication and authorization; idle-mode UE tracking and reachability; security negotiations; and NAS (Non Access Stratum)† signaling. It is connected to EUTRAN via S1-MME interface.

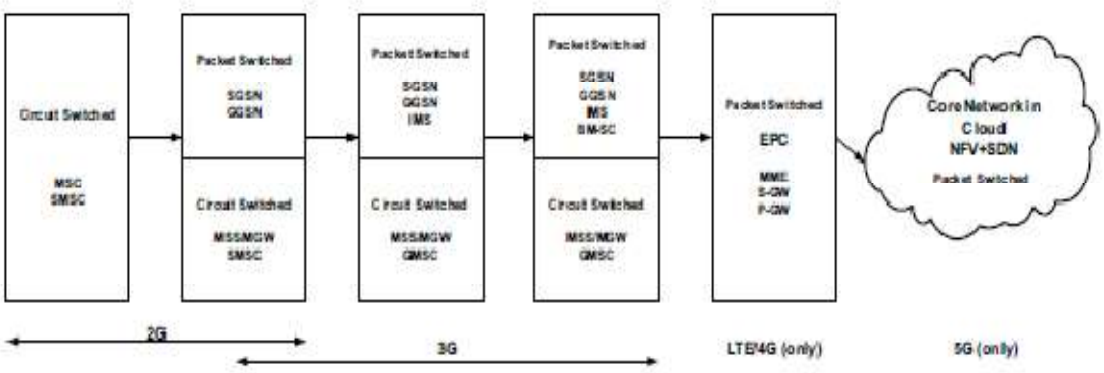


Figure 1.7: Core Network Evolution

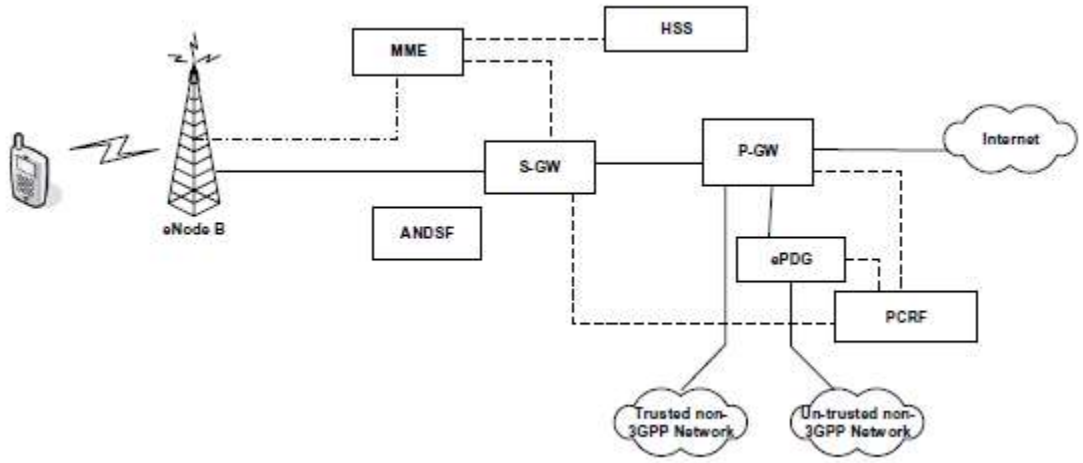


Figure 1.8: Evolved Packet Core

The S-GW connects to EUTRAN via S1-U interface. For each UE associated with the EPS, at a given point of time, there is one single S-GW. The S-GW routes and forwards user data packets, while also acting as the mobility anchor for the user plane during inter-eNodeB handovers and as the anchor for mobility between inter-3GPPP technologies. For idle state UEs, the S-GW terminates the downlink data path and triggers paging when downlink data arrives for the UE. It manages and stores UE contexts, for example, parameters of the IP bearer service and network internal routing information. It also performs replication of the user traffic in case of lawful interception.

The Packet Gateway (PGW) serves as the entering point in EPS (Evolved Packet System). It is connected to external PDNs, operators' IMS, and non-IMS IP services, and provides access for trusted and nontrusted non-3GPP IP networks. It acts as the anchor for mobility between 3GPP and non-3GPP access systems. The PGW performs policy enforcement, packet filtering (by, e.g., deep packet inspection*) for each user, charging support, and lawful interception. It also performs transport level packet marketing in the uplink and downlink.

EPC also consists of the following elements to support authentication and mobility management.

The HSS (Home Subscriber Server) is a central database that contains user related and subscription related information. The functions of the HSS include functionalities such as mobility management, call and session establishment support, user authentication, and access authorization. The HSS is based on pre-Rel-4 Home Location Register (HLR) and Authentication Center (AuC).

The ANDSF (Access Network Discovery and Selection Function) provides information to the UE about connectivity to other 3GPP and also to non-3GPP access networks (such as WiFi). The purpose of the ANDSF is to assist the UE to discover the access networks in their vicinity and to provide rules (policies) to prioritize and manage connections to these networks.

The ePDG's (Evolved Packet Data Gateway) main function is to secure the data transmission with a UE connected to the EPC over an untrusted non-3GPP access. For this purpose, the ePDG acts as a termination node of IPsec* tunnels established with the UE.

The Policy and Charging Rules Function (PCRF) is an optional component which can be supported even when IMS is not supported. It determines policy rules in a multimedia network in real time. It also supports service data flow detection, policy enforcement, and flow-based charging.

Core networks traditionally have been designed as a single architecture addressing a range of requirements and supporting backward compatibility. This one size fits all approach has been successful in keeping the costs down to a reasonable level and by supporting legacy circuit switched and today's packet switched functionalities. This core network, however, is rigid in the sense that it is not flexible enough to accommodate the customized and variable connectivity needs of individual users and businesses that are expected in the future. However, with virtualization, NFV, SDN, and network slicing, it is possible to make core networks more flexible and scalable. Thus, the next generation core network is expected to exist in a cloud-based environment with a high degree of virtualization and software-based networking. Such flexibility is needed to support a variety of access networks such as 3G, LTE, 4G, WiFi, and tomorrow's 5G.

1.6.2 Components of Core Network:

The current EPC will further evolve to support virtualization and network slicing to become NGC applicable for 5G networks. Network slicing is often termed as logical instantiation of a network possibly due to virtualization technologies. The concept is seen as the natural extension/evolution of the current network sharing methodologies.

Network slicing is one of the promising techniques that will likely exist in both radio access and core networks. It allows multiple logical networks to be created on top of a common physical infrastructure. Either DCN (Dedicated Core Network) or a combination of NFV and SDN can be used as a technology to enable network slicing along with orchestration and analytics. DCN or Décor as defined in 3GPP TS 23.401 is a feature that enables an operator to deploy multiple logical mobile core networks connected to the same RAT or multiple RATs

(e.g. GERAN, UTRAN, E-UTRAN, WB-E-UTRAN and NB-IoT). A DCN consists of one or more MME/SGSN and it may be comprised of one or more SGW/PGW/PCRF.

This feature enables subscribers to be allocated to and served by a DCN based on subscription information (e.g., "UE Usage Type"). With 5G, a single terminal can use multiple services with different characteristics almost simultaneously. In such cases, a network slice can be created for each service, requiring all such slices to coordinate control for that particular single terminal. These slices can be mapped to respective radio and core network slices to provide end-to-end connectivity. The methodology is currently being specified for selecting radio/core networks particularities for supporting slicing in existing as well as in future 5G systems. Control and User Planes' Separation: The separation of control and user planes is one of the key principles of 5G core network architecture. 3GPP started a study in TR 23.714 on user/control planes' separation involving core network elements such as P-GW, Traffic Detection Function (TDF), and so on.

This separation allows independent scaling of each plane and migration toward cloud-based architecture. For example, the control plane can be placed in a centralized location with complex hardware and processing capabilities. On the other hand, the user plane can be distributed to a larger number of local sites making reachability from the perspective of a user easier. A good example of this will be content caching in local sites instead of securing it from the main server sitting thousands of miles away. This separation is the fundamental concept of SDN and having SDN will make core networks more flexible.

1.7 5G Standardization:

5G Protocol standardization is the process of tailoring the 5G technology to serve the market requirements and even more, by introducing new applications and services besides the traditional services introduced by the initial mobile networks such as 1G, 2G, 3G, and 4G. The 5G standardization process has been a responsibility of OTSA (Open Trial Specification Alliance), which was tasked to accelerate the standardization and commercial deployment process of 5G. The standardization process adopted new 5G technologies such as New Radio (NR) and NextGen Core, while some stemmed from the initial mobile technologies such as Long Term Evolution (LTE).

1.7.1 3GPP (3rd Generation Partnership Project) Releases: 3GPP comprises of several Releases, among which are; R99. Rel-4 to Rel-16. The newly proposed Rel-17 and Rel-18. Rel-14 and all the initial releases define the previous mobile networks such as 4G, 3G, 2G, and 1G. LTE is defined from Rel-8, LTE-A is from Rel-10, and LTE-A Pro is from Rel-12, to mention just but few recent releases.

5G is defined in 3GPP Release 15 (Rel-15) and Release 16 (Rel-16), which constitute the following: NextGen Core (NGC) network. New Radio (NR). LTE Advanced Pro Evolution. EPC Evolution. Among the 5G, new technologies are New Radio and NextGen Core network. Other technologies that have been improved from some of the preceding Releases of 3GPP are EPC Evolution and LTE Advanced pro Evolution.

1.7.2 Releases of 5G:

Rel-15: It is popularly considered the basic version of 5G. It is phase 1 of the 5G system that implemented the following improvement on NR:

Construct the NR technical framework:

- New waveform - the F-OFDM technology is used.
- Coding modulation and channel. Massive MIMO (Multiple Input Multiple Output) - supports up to 64T64R.
- Numerology, frame structure - refers to the change of the timeslot length and frame structure caused by different subcarrier spacing.
- Flexible duplex - the uplink and downlink configurations are flexible. Also, the uplink and downlink can be included in the same timeslot.

Network architecture ready:

- Non-standalone/Standalone.
- Uplink and downlink decoupling.
- CU-DU high-level segmentation.

1.7.3 Development of uRLLC and mMTC:

The improvement in Rel-15 functions to Rel-16 provides a complete uRLLc low latency and highly reliable capabilities. URLLC service explores the industry's network requirements and further improves standards, technologies, and deployment specifications. Some major applications of this advantage are the Internet of Things (IoT), virtual reality, Augmented Reality, Mixed Reality, and many more. In mMTC, 5G will coexist with NB/eMTC, which may be improved in the future NR system. Other applications, such as NB IoT/eMTC technology, are still evolving.

1.8 5G Network security:

The goal of 5G network security is to protect user data and enable network resilience and business continuity. To ensure this, 5G has designed security measures that address many of the threats faced in today's 4G/3G/2G networks and meet network security demands.

Some of the demands on network security are as follows:

- Availability: The identification of illegal attacks and reduction of their impact.
- Traceability: Recording of operations for security audit and problem identification.
- Integrity and Confidentiality: Protection of user privacy information, user communication data, and operator's principal data.

The security measures try to ensure the above by implementing the following:

- Enhanced security.

- Stronger security on the air interface; the user plane has integrity protection by anti-alter, unlike 4G that is prone to user plane attack.
- User privacy protection such that the users' IMSI is encrypted, unlike 4G, transmits user IMSI in plaintext.
- Improved interconnection security by implementing end-to-end protection between PLMNs, unlike 4G that is similar to SS7 attacks.
- Improved cryptographic algorithm using a 256-bit cryptographic algorithm vis-a-vis 4G that uses a 128-bit cryptographic algorithm.

1.9 5G Deployment:

The keys to successful 5G network deployment adapt traditional best practices to the new technological breakthroughs that have set 5G apart. These principles cut across all facets of 5G network architecture, technology, and performance.

- Certify all fiber connections and validate orientation/alignment of antenna: The importance of good fiber hygiene has been magnified by the manifold increase in antenna connections inherent to 5G Massive MIMO. The commitment to high-quality connection and validation must also extend to coax installation for the FR1 band. Exceeding the link budget can lead directly to performance degradation and delayed turn-on. Antenna alignment, including both orientation and tilt verification, provides a valuable baseline for optimized 5G cell site performance and coverage.
- Verify carrier and SSB spacing frequency, and subcarrier spacing: The synchronization signaling block (SSB), which is the 5G equivalent of the LTE reference signal, is used to identify and synchronize a cell with specific user equipment (UE). Each SSB can be identified by a unique number known as the SSB index. A UE will latch onto a specific beam, based on the SSB index with the highest observed signal strength. Verifying SSB functionality is critical during 5G network deployment and commissioning. The performance and spacing of each subcarrier should also be tested.
- Verify all carriers are present and PCI of each carrier: A robust 5G network deployment plan should include signal verification for each carrier and their respective physical cell ID (PCI). Carrier aggregation is a technique used to increase the data rate per user by assigning multiple frequency blocks or component carriers to each. Improved utilization through carrier aggregation is an important enabler of 5G bandwidth and use case diversity.
- Verify beam IDs for each carrier: In an LTE deployment, coverage can be blanketly characterized by sector. Using 5G NR technology, each individual beam behaves much like a separate coverage area all its own. The "beam-centric" philosophy of 5G underscores the importance of dedicated beam index analysis as part of 5G network deployment.
- Verify 5G site coverage: Verifying the cell coverage output designed into the 5G network requires accurate 5G coverage mapping to determine beam index, power, and signal-to-noise ratio for a given area. This 5G deployment best practice can be

difficult to achieve reliably, particularly for combined 5G and LTE coverage areas. Dynamic Spectrum Sharing (DSS) enables 5G and LTE to operate in tandem for seamless coverage and rapid 5G deployment. The best 5G coverage mapping tools are now provisioned for concurrent LTE and 5G coverage assessment.

1.9.1 5G Deployment Challenges:

With so many options to choose from, simply deciding which fifth generation approach to take is the first of many inherent deployment challenges. Breakthrough 5G wireless technology platforms are pushing the envelope of design, manufacturing, and testing capabilities. Network Function Virtualization (NFV) is a prerequisite for core network slicing, intelligence at the edge, and other essential 5G signal features. These technologies power the delivery of IoT and AI-based services. Standardization, security, and the requisite CPU horsepower to drive virtual functions are some of the many obstacles being tackled by NFV developers.

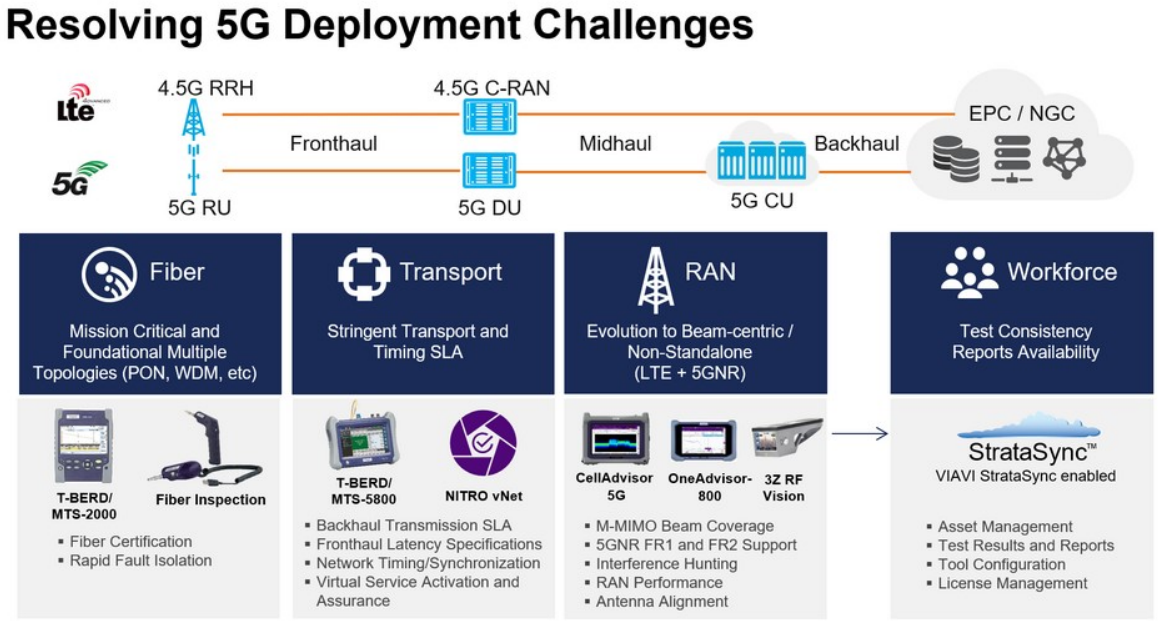


Figure 1.9: Resolving 5G Deployment Challenges

The millimeter wave is another essential fifth generation ingredient that can present technological and logistical challenges. Due to the limited range and inability to transmit through solid objects, the sheer volume of antennae required introduces hurdles that can only be addressed through methodical, incremental deployment. Spectral efficiency, measured in (bit/s)/Hz, is currently gated by the Shannon Limit which defines the maximum rate that data can be sent over any medium with zero error. This theoretical ceiling is much less than what is expected and required for 5G deployment. Only Massive MIMO and beamforming, utilizing large antenna arrays, enables 5G to effectively circumvent this natural limit of faster speeds.

1.10 Ten Pillars of 5G:

We identify 10 key building blocks for 5G, illustrated by Figure 1.3. In the following, we elaborate each of these blocks and highlight their role and importance for achieving 5G.

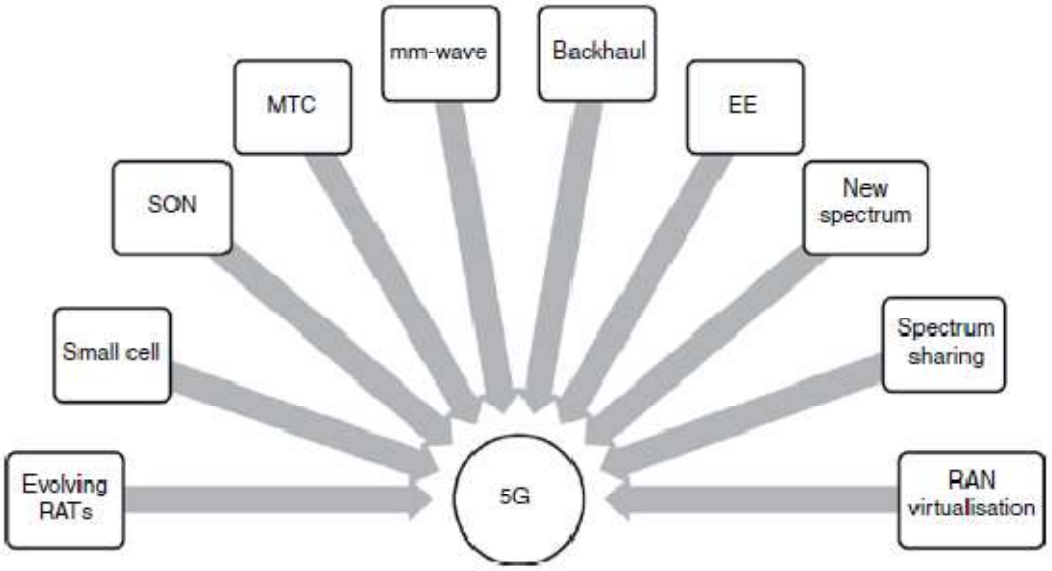


Figure 1.10: Ten Pillars of 5G

1.10.1 Evolution of Existing RATs

As mentioned before, 5G will hardly be a specific RAT, rather it is likely that it will be a collection of RATs including the evolution of the existing ones complemented with novel revolutionary designs. As such, the first and the most economical solution to address the 1000x capacity crunch is the improvement of the existing RATs in terms of SE, EE and latency, as well as supporting flexible RAN sharing among multiple vendors. Specifically, LTE needs to evolve to support massive/3D MIMO to further exploit the spatial degree of freedom (DOF) through advanced multi-user beamforming, to further enhance interference cancellation and interference coordination capabilities in a hyperdense small-cell deployment scenario. WiFi also needs to evolve to better exploit the available unlicensed spectrum. IEEE 802.11ac, the latest evolution of the WiFi technology, can provide broadband wireless pipes with multi-Gbps data rates. It uses wider bandwidth of up to 160 MHz at the less polluted 5 GHz ISM band, employing up to 256 Quadrature Amplitude Modulation (QAM). It can also support simultaneous transmissions up to four streams using multi-user MIMO technique. The incorporated beamforming technique has boosted the coverage by several orders of magnitude, compared to its predecessor (IEEE 802.11n). Finally, major telecom companies such as Qualcomm have recently been working on developing LTE in the unlicensed spectrum as well as integrating 3G/4G/WiFi transceivers into a single multi-mode base station (BS) unit. In this regard, it is envisioned that the future UE will be intelligent enough to select the best interface to connect to the RAN based on the QoS requirements of the running application.

1.10.2 Hyperdense Small-Cell Deployment:

Hyperdense small-cell deployment is another promising solution to meet the 1000x capacity crunch, while bringing additional EE to the system as well. This innovative solution, also referred to as HetNet, can help to significantly enhance the area spectral efficiency (b/s/Hz/m2). In general, there are two different ways to realise HetNet: (i) overlaying a cellular system with small cells of the same technology, that is, with micro-, pico-, or femtocells; (ii) overlaying with small cells of different technologies in contrast to just the cellular one (e.g. High Speed Packet Access (HSPA), LTE, WiFi, and so on).

The former is called multi-tier HetNet, while the latter is referred to as multi-RAT HetNet. Qualcomm, a leading company in addressing 1000x capacity challenge through hyperdense small-cell deployments, has demonstrated that adding small cells can scale the capacity of the network almost in a linear fashion. That is, the capacity doubles every time we double the number of small cells. However, reducing the cell size increases the inter-cell interference and the required control signalling. To overcome this drawback, advanced inter-cell interference management techniques are needed at the system level along with complementary interference cancellation techniques at the UEs. Small-cell enhancement was the focal point of LTE R-12, where the New Carrier Type (NCT) (also known as the Lean Carrier) was introduced to assist small cells by the host macro-cell. This allows more efficient control plane functioning (e.g. for mobility management, synchronisation, resource allocation, etc.) through the macro-layer while providing a high-capacity and spectrally efficient data plane through the small cells. Finally, reducing the cell size can also improve the EE of the network by bringing the network closer to the UEs and hence shrinking the power budget of the wireless links.

1.10.3 Self-Organising Network

Self-Organising Network (SON) capability is another key component of 5G. As the population of the small cells increases, SON gains more momentum. Almost 80% of the wireless traffic is generated indoors. To carry this huge traffic, we need hyperdense small-cell deployments in homes – installed and maintained mainly by the users – out of the control of the operators. These indoor small cells need to be self-configurable and installed in a plug and play manner. Furthermore, they need to have SON capability to intelligently adapt themselves to the neighbouring small cells to minimise inter-cell interference. For example, a small cell can do this by autonomously synchronising with the network and cleverly adjusting its radio coverage.

1.10.4 Machine Type Communication

Apart from people, connecting mobile machines is another fundamental aspect of 5G. Machine type communication (MTC) is an emerging application where either one or both of the end users of the communication session involve machines. MTC imposes two main challenges on the network. First, the number of devices that need to be connected is tremendously large. Ericsson (one of the leading companies in exploring 5G) foresees that 50 billion devices need to be connected in the future networked society; the company envisages ‘anything that can benefit from being connected will be connected’. The other challenge imposed by MTC is the accelerating demand for real-time and remote control of mobile devices (such as vehicles)

through the network. This requires an extremely low latency of less than a millisecond, socalled "tactile Internet", dictating 20x latency improvement from 4G to 5G.

1.10.5 Developing Millimetre-Wave RATs

The traditional sub-3 GHz spectrum is becoming increasingly congested and the present RATs are approaching Shannon's capacity limit. As such, research on exploring cm- and mmWave bands for mobile communications has already been started. Although the research on this field is still in its infancy, the results look promising. There are three main impediments for mmWave mobile communications.

First, the path loss is relatively higher at these bands, compared to the conventional sub-3GHz bands. Second, electromagnetic waves tend to propagate in the Line-Of-Sight (LOS) direction, rendering the radio links vulnerable to being blocked by moving objects or people. Last but not least, the penetration loss through the buildings is substantially higher at these bands, blocking the outdoor RATs for the indoor users. Despite these limitations, there are myriad advantages for mmWave communications.

An enormous amount of spectrum is available in mmWave band; for example, at 60 GHz, there is 9GHz of unlicensed spectrum available. This amount of spectrum is huge, especially when we think that the global allocated spectrum for all cellular technologies hardly exceeds 780 MHz. This amount of spectrum can completely revolutionise mobile communications by providing ultra-broadband wireless pipes that can seamlessly glue the wired and the wireless networks. Other advantages of mmWave communications include the small antenna sizes ($\lambda/2$) and their small separations (also around $\lambda/2$), enabling tens of antenna elements to be packed in just one square centimetre. This in turn allows us to achieve very high beamforming gains in relatively small areas, which can be implemented at both the BS and the UE. Incorporating smart phased array antennas, we can fully exploit the spatial degree of freedom of the wireless channel (using Space-Division Multiple Access (SDMA)), which can further improve the system capacity. Finally, as the mobile station moves around, beamforming weights can be adjusted adaptively so that the antenna beam is always pointing to the BS.

1.10.6 Redesigning Backhaul Links

Redesigning the backhaul links is the next critical issue of 5G. In parallel to improving the RAN, backhaul links also need to be reengineered to carry the tremendous amount of user traffic generated in the cells. Otherwise, the backhaul links will soon become bottlenecks, threatening the proper operation of the whole system. The problem gains more momentum as the population of small cells increases. Different communication mediums can be considered, including optical fibre, microwave and mmWave. In particular, mmWave point-to-point links exploiting array antennas with very sharp beams can be considered for reliable self-backhauling without interfering with other cells or with the access links.

1.10.7 Energy Efficiency:

EE will remain an important design issue while developing 5G. Today, Information and Communication Technology (ICT) consumes as much as 5% of the electricity produced around

the globe and is responsible for approximately 2% of global greenhouse gas emissions – roughly equivalent to the emissions created by the aviation industry. What concerns more is the fact that if we do not take any measure to reduce the carbon emissions, the contribution is projected to double by 2020. Hence, it is necessary to pursue energy-efficient design approaches from RAN and backhaul links to the UEs.

The benefit of energy-efficient system design is manifold. First, it can play an important role in sustainable development by reducing the carbon footprint of the mobile industry itself. Second, ICT as the core enabling technology of the future smart cities can also play a fundamental role in reducing the carbon footprint of other sectors (e.g. transportation). Third, it can increase the revenue of mobile operators by reducing their operational expenditure (Opex) through saving on their electricity bills. Fourth, reducing the 'Joule per bit' cost can keep mobile services affordable for the users, allowing flat rate pricing in spite of the 10 to 100x data rate improvement expected by 2020. Last but not least, it can extend the battery life of the UEs, which has been identified by the market research company TNS as the number one criterion of the majority of the consumers purchasing a mobile phone.

1.10.8 Allocation of New Spectrum for 5G:

Another critical issue of 5G is the allocation of new spectrum to fuel wireless communications in the next decade. The 1000x traffic surge can hardly be managed by only improving the spectral efficiency or by hyper-densification. In fact, the leading telecom companies such as Qualcomm and NSN believe that apart from technology innovations, 10 times more spectrum is needed to meet the demand. The allocation of around 100 MHz bandwidth at the 700 MHz band and another 400 MHz bandwidth at around 3.6 GHz, as well as the potential allocation of several GHz bandwidths in cm- or mmWave bands to 5G will be the focal point of the next WRC conference, organised by ITU-R in 2015.

1.10.9 Spectrum Sharing:

Regulatory process for new spectrum allocation is often very time consuming, so the efficient use of available spectrum is always of critical importance. Innovative spectrum allocation models (different from the traditional licensed or unlicensed allocation) can be adopted to overcome the existing regulatory limitations. Plenty of radio spectrum has traditionally been allocated for military radars where the spectrum is not fully utilised all the time (24/7) or in the entire geographic region. On the other hand, spectrum cleaning is very difficult as some spectrum can never be cleaned or can only be cleaned over a very long time; beyond that, the spectrum can be cleaned in some places but not in the entire nation. As such, the Authorised/Licensed Shared Access (ASA/ LSA) model has been proposed by Qualcomm to exploit the spectrum in small cells (with limited coverage) without interfering with the incumbent user (e.g. military radars). This kind of spectrum allocation model can compensate the very slow process of spectrum cleaning. It is also worth mentioning that as mobile traffic growth accelerates, spectrum refarming becomes important, to clean a previously allocated spectrum and make it available for 5G. Cognitive Radio concepts can also be revisited to jointly utilise

licensed and unlicensed spectrums. Finally, new spectrum sharing models might be needed as multi-tenant network operation becomes widespread.

1.10.10 RAN Virtualisation:

The last but not least critical enabler of 5G is the virtualisation of the RAN, allowing sharing of wireless infrastructure among multiple operators. Network virtualisation needs to be pushed from the wired core network (e.g. switches and routers) towards the RAN. For network virtualisation, the intelligence needs to be taken out of the RAN hardware and controlled in a centralised manner using a software brain, which can be done in different network layers. Network virtualisation can bring myriad advantages to the wireless domain, including both Capex (Capital Expenditure) and Opex savings through multi-tenant network and equipment sharing, improved EE, on-demand up- or down-scaling of the required resources, and increased network agility through the reduction of the time-to-the-market for innovative services (from 90 hours to 90 minutes), as well as easy maintenance and fast troubleshooting through increased transparency of the network [14]. Virtualisation can also serve to converge the wired and the wireless networks by jointly managing the whole network from a central orchestration unit, further enhancing the efficiency of the network. Finally, multi-mode RANs supporting 3G, 4G or WiFi can be adopted where different radio interfaces can be turned on or off through the central software control unit to improve the EE or the Quality of Experience (QoE) for the end users.

Reference Books:

1. Saad Z. Asif, "5G Mobile Communications Concepts and Technologies, CRC Press, 1st Edition, 2019.
2. Erik Dahlman, Stefan Parkvall, Johan Skold "5G NR: The Next Generation Wireless Access Technology", Academic Press, 1st Edition, 2018.
3. Jonathan Rodriguez, "Fundamentals 5G Mobile Networks", John Wiley & Sons, 1st Edition, 2015.
4. Long Zhao, Hui Zhao, Kan Zheng, Wei Xiang, "Massive MIMO in 5G Networks: Selected Applications", Springer, 1st Edition, 2018.
5. Robert W. Heath Jr., Angel Lozano, "Foundations of MIMO Communication", Cambridge University Press, 1st Edition, 2019.
6. R. Vannithamby and S. Talwar, "Towards 5G: Applications, Requirements and Candidate Technologies", John Willey & Sons, 1st Edition, 2017.

Questions to Practice:

PART -A

1 Identify how 5G technology is going to be a trendsetter in future

2 Identify the body which showcase general policy, strategy and perform various tasks for any new cellular Technology.

3 Describe in detail about 5G NR

4 Classify the use cases of 5g that will impact the way data is transmitted over cellular networks.

5 Explain in detail about Spectral Efficiency

PART-B

1 With the advent of 5G, Demonstrate how industry experts define the New Radio and Core network that should evolve to support the needs of 5G.

2 Classify how Spectrum and Deployment of 5G is used in current scenario

3 Demonstrate in detail about the Evolving LTE to 5G Capability

4 Demonstrate in detail about the impact of Self Organising Network, Machine Learning, mmWave, Backhaul in 5G Communication

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF ELECTRICAL AND ELECTRONICS

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

UNIT – II – 5G CHANNEL ACCESS METHODS– SECA3020

UNIT-II (5G CHANNEL ACCESS METHODS)

OFDM and OFDMA – MIMO OFDM – Generalized Frequency Division Multiplexing (GFDM) – Non-Orthogonal Multiple Access (NOMA) - Universal Filtered OFDM -Filter bank multicarrier (FBMC)- Sparse Code Multiple Access (SCMA) –Comparison of multiple access methods

2.1 Introduction :

One of most attractive traits of mobile communications is the wireless connectivity between the device and the network. Having a perfect mobile wireless connection for voice, video, and data indulgence is the need for today and tomorrow. However, this wireless connectivity presents one of most daunting and fundamental challenges of this field.

Wireless connectivity in mobile communications is directly associated with frequency spectrum and channel access method (multiple access method). Spectrum is scarce and expensive but is a must-have to run a mobile network. The channel (atmosphere in this case) is unpredictable and beyond anyone's control. The daunting challenge is to squeeze in more bps/Hz/km^2 (bits per second per hertz per square kilometer), which is called system spectral efficiency. To address this challenge, beside other countless innovations, almost every generation of mobile communication has come up with a new multiple access method as an improvement over the previous one. The 2G systems use FDMA (Frequency Division Multiple Access) and TDMA (Time Division Multiple Access) techniques. The 3G systems use CDMA (Code Division Multiple Access) while LTE, WiMAX, and 4G systems employ OFDMA (Orthogonal FDMA). Thus, it can be safely said and with almost certainty that 5G will be embedded with one or more new and improved multiple access method(s) or implanted with an existing one with sufficient improvements.

Varied methods of orthogonal and nonorthogonal multiple accesses for 5G systems have been under research and investigation for at least the last five years. Nonorthogonal multiple access is more suitable for uplink since the base station can afford the multiuser detection complexity. On the other hand, for downlink, orthogonal multiple access is more suitable due to the limited processing power of the user equipment [3,4]. The focus of this section is on some of the developing channel access methods which are under consideration for 5G systems. However, before diving into the details of such potential methods, it may be productive to refresh basic understandings of channel capacity and spectral efficiency.

2.2 Fundamental Concepts:

Air-Interface: The air-interface defines the method for transmitting/receiving information over the air between mobiles and base stations. The air interfaces of 2G, 3G, and 4G were all designed while keeping certain KPIs (Key Performance Indicators) in mind (for example, mean opinion score for voice, dropped/blocked call rates, data throughput, etc.). However, the emerging trends of IoTs, M2M (Machine to Machine), V2X, and so on are all demanding to go beyond such a static/specific approach. Channel Capacity: Communicating messages from one location to another requires some form of pathway or medium. The communications channel is

any medium (wired or wireless) over which information can be transmitted/received. Cellular communications use radio waves to carry information over the air from the user to the base station and vice versa. Channel capacity is the tight upper bound on the rate at which information can be transmitted with an arbitrarily small error probability over a communications channel. The famous Shannon–Hartley theorem provides this channel capacity as elaborated in Equation 2.1

$$C=B*\log_{2*}[(1+(S/N)]\text{----------Equation 2.1}$$

where

C is the channel capacity in bits per second

B is the bandwidth of the channel in Hertz

S is the average received signal power over the bandwidth, measured in watts

N is the average noise or interference power over the bandwidth, measured in watts (or volts squared)

Figure 2.1: Channel Access Methods

In a point to point case, if R (actual bit rate in bps) ≤ C, the theorem provides the maximum rate at which information can be transmitted over a communications channel under a specified bandwidth in the presence of noise.. If R > C, then errorless communication is next to impossible. When there is more than one user, that is, in a multiuser case, the concept may be extended to a set of all pairs (R1, R2) such that both user 1 and user 2 can simultaneously communicate at rates R1 and R2, respectively. Under this scenario, when the bandwidth is shared, one user may communicate at a higher rate and the other at a lower rate. For example, in OFDM, this tradeoff is achieved by varying the number of subcarriers allocated to each user.

Channel Access Methods: A channel access method is based on multiplexing allowing sharing of a communication channel between users/devices. This form of multiplexing is based on the physical layer or layer 1 of the OSI (Open Systems Interconnection) model. A channel access method can also be based on media access control (MAC) which is the sublayer of layer 2

(Data Link Layer) of the OSI model. This section will focus on the channel access methods applicable to the physical layer.

The three multiple access techniques that are currently prevailing in mobile communications are FDMA, TDMA, and CDMA as shown in Figure 2.1. FDMA provides different frequency bands to different data streams whereas TDMA provides different time slots to different data streams. In CDMA, several message signals are transferred simultaneously over the same carrier frequency, utilizing different spreading codes. OFDMA, which is used in 4G standards, is a form of FDMA. OFDM achieves high spectral efficiency by using orthogonal subcarriers. Orthogonality allows subcarriers' spectra to overlap, which in turn, enables transmission of more data than FDMA over the same fixed bandwidth.

However, OFDM does have drawbacks such as the spectrum is not localized and requires a guard band. The subcarrier spacing and symbol duration are fixed and transmission is synchronous mandating a large overhead for time alignment. These shortcomings make OFDM less attractive for some usage scenarios of 5G.

Spectral Efficiency: The spectral efficiency refers to the information rate that can be transmitted over a given bandwidth in a specific communication system. The spectral efficiency of a mobile communications system largely depends on the choice of a multiple access method. The other factors may include the type of modulation used, error correction methods, frequency reuse factor, the number of users served, radio capability, and the percentage of time a service is active. However, spectral efficiency of a technology is largely independent of the frequency at which it operates, since modulation and coding are the same at different frequencies.

This could be measured as bit/s/Hz which is called the link spectral efficiency or bit/s/Hz per cell (site) which is system spectral efficiency. The system spectral efficiency is more practical as clarifies how efficiently an operator has deployed a specific amount of spectrum. The U.S. FCC TAC (Technology Advisory Council) specifically recommends bps/Hz/km2 as the metric for Personal Communications Systems,* which takes into account both spectral efficiency (bps/Hz) and deployment density. Today's advanced wireless technologies are essentially close to reaching the Shannon Bound, which defines the maximum/upper theoretical efficiency possible relative to noise. Thus, future gains in spectral efficiency will be limited unless a better multiple access technique is developed or sufficient improvement is made over an existing one for 5G networks.

2.3 Multiple Access Waveform:

A one-size-fits-all air-interface, which has been the typical solution for the past twenty plus years, may no longer be the total solution for 5G. First, since the available spectrum bands for 5G can be distributed over a large range of frequencies, including even the millimeter wave bands, the air-interface should be flexible enough so that it can operate in different frequency bands. Together with advanced radio frequency (RF) architecture and RF-related signal processing, it needs to support either flexible switching between different frequency bands or simultaneous operation in several frequency bands, including fragmented usage of certain bands. For this purpose, flexible numerology and frame structure as well as adaptive configuration are

needed . 3GPP through its Rel-14 has endorsed OFDM-based waveform for eMBB operating up to 40 GHz in downlink and uplink. DFT-S-OFDM (Discrete Fourier transform Spread OFDM) based and CP-OFDM (Cyclic Prefix OFDM) waveforms are also supported in uplink for eMBB operating up to 40 GHz [13]. The radio characteristics for URLLC service are getting defined in Rel-15 and Rel-16 [14]. OFDM avoids interference and creates a high capacity but requires a lot of signaling and increases delay. Delay may not be suitable for URLLC and heavy signaling for mMTC types of applications . Therefore, to optimize a wide variety of 5G services, a number of waveforms including OFDM were discussed in 3GPP. The candidate multiple access methods can be characterized by signatures (attributes) such as use of codebooks, orthogonal/nonorthogonality mode, and the presence of an interleaver/scrambler. At the receiver, multi-user detection schemes are employed to extract the original data on a per user basis. A high-level description of these access methods is discussed in this section.

2.3.1 OFDM:

OFDM is a multi-carrier modulation technique developed in the 1960s. The first OFDM-based standard was Digital Audio Broadcasting developed by ETSI in 1995. Since then, OFDM has been part and parcel of many telecom/broadcasting standards and its CP-OFDM form is currently used in LTE, WiMAX, and LTE-Advanced (4G) standards. OFDM capitalizes on the use of cyclic prefixes to reduce intersymbol interference (ISI) and IFFT/FFT operations. IFFT/FFT (Inverse/Fast Fourier transform) allow combining multiple carriers at the baseband leading to OFDMA. OFDMA offers bandwidth scalability, robustness to multipaths, and effective integration with MIMO. However, aside from the benefits, OFDM suffers from high PAPR (Peak-to-Average Power Ratio) and inferior frequency localization due to the use of pulse shape filters. Details on OFDM/OFDMA can be found. To overcome such limitations, some add-ons may be added to OFDM and with these additional attributes it may become a suitable access method for 5G. Add-ons such as Weighted Overlap and Add (WOLA) ,which replaces the rectangular pulse with a pulse with soft edges at both sides, result in much sharper sidelobe decay in the frequency domain. This decay reduces the out-of-band (OOB) leakage at the transmitter end. At the receiver, WOLA provides suppression of other (asynchronous) users' interference.

In OFDM systems, only a single user can transmit on all of the sub-carriers at any given time. In order to support multiple users time and/or frequency division access techniques are used in OFDM. The major setback to this static multiple access scheme is the fact that the different users see the wireless channel differently is not being utilized. OFDMA, on the other hand, allows multiple users to transmit simultaneously on the different sub-carriers per OFDM symbol. OFDM is employed in Fixed WiMAX system deployed around the world for broadband internet service. Figure 2.2 depicts OFDM frame structure employed in fixed WiMAX system. Here Downlink sub frame is transmitted by Base station to subscriber stations and Uplink sub frame is transmitted by multiple subscriber stations to the Base Station. Both the frame is composed of more than one OFDM symbols and each symbol is made up of subcarriers, which fall in data and pilot subcarriers, where data subcarriers carry the user data. There are 192 data sub carriers in Fixed WiMAX System. The point here is Subscriber station

has been assigned one or more symbols by BS and all the data carriers(i.e. 192) of the symbols are occupied by one SS. It is depicted in the figure-1 that entire 256 carriers are allocated to the one user statically in TDD frame.

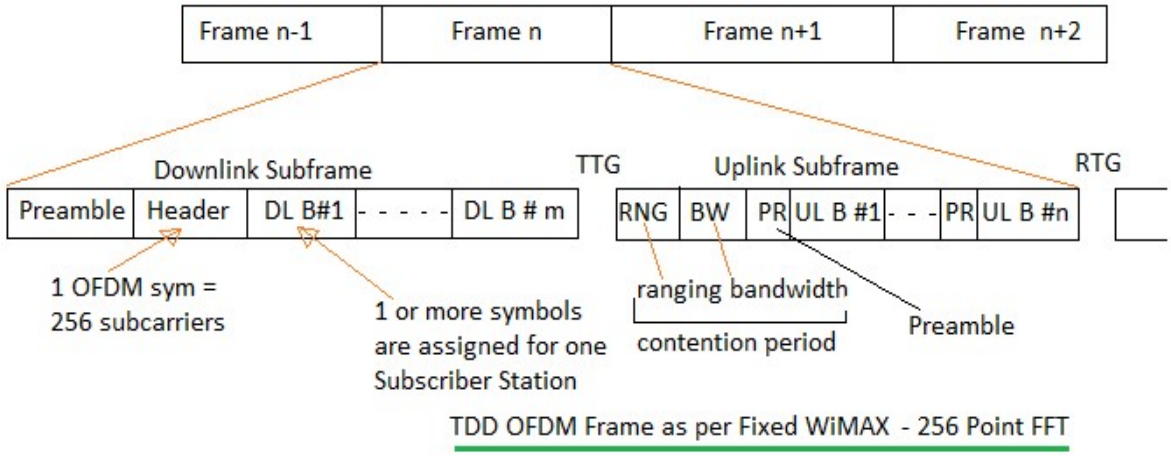


Figure 2.2 OFDM modulation frame structure as per 16d standard

To understand difference between OFDM and OFDMA, one should understand basic difference between OFDM and FDM multiplexing techniques in addition to OFDM physical layer and OFDMA physical layer as per fixed wimax and mobile wimax standards.

2.3.2 OFDMA:

OFDM is a multi-carrier modulation technique developed in the 1960s. The first OFDM-based standard was Digital Audio Broadcasting developed by ETSI in 1995. Since then, OFDM has been part and parcel of many telecom/broadcasting standards and its CP-OFDM form is currently used in LTE, WiMAX, and LTE-Advanced (4G) standards. OFDM capitalizes on the use of cyclic prefixes to reduce intersymbol interference (ISI) and IFFT/FFT operations. IFFT/FFT (Inverse/Fast Fourier transform) allow combining multiple carriers at the baseband leading to OFDMA. OFDMA offers bandwidth scalability, robustness to multipaths, and effective integration with MIMO. However, aside from the benefits, OFDM suffers from high PAPR (Peak-to-Average Power Ratio) and inferior frequency localization due to the use of pulse shape filters. Details on OFDM/OFDMA can be found. To overcome such limitations, some add-ons may be added to OFDM and with these additional attributes it may become a suitable access method for 5G. Add-ons such as Weighted Overlap and Add (WOLA) , which replaces the rectangular pulse with a pulse with soft edges at both sides, result in much sharper sidelobe decay in the frequency domain. This decay reduces the out-of-band (OOB) leakage at the transmitter end. At the receiver, WOLA provides suppression of other (asynchronous) users' interference.

In the case of OFDMA, which is employed in Mobile WiMAX system deployed around the world and also employed in LTE system being deployed, total subcarriers are permuted and assigned to sub channel. Hence many SSs can occupy the same sub channel but use different subcarriers to transmit the information. Figure 2.3 describes OFDMA frame used in Mobile

WiMAX System. It clearly mentions that one symbol is composed of more than one sub channel and each sub channel is composed of distributed subcarriers. The point here is each symbol is used by more number of SSs to transmit and receive the information which is depicted by Burst 1 and Burst 2 in the figure. As mentioned in OFDMA subcarriers are divided among users at the same time instant. Figure mentions 2048 FFT case here. Total 2048 subcarriers of FFT is divided among 60 subchannels. Each subchannels will have their own pilot and data subcarriers.

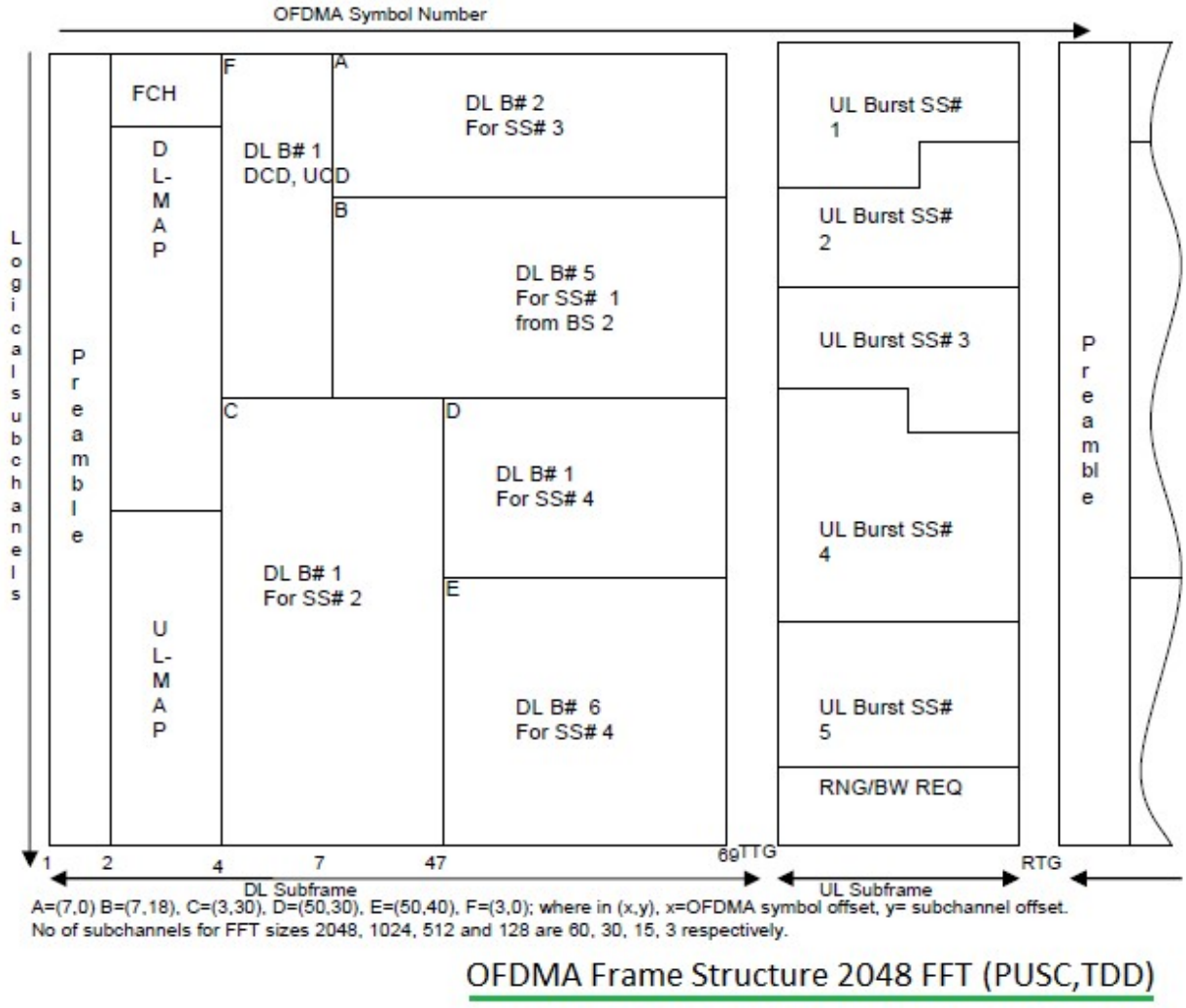


Figure 2.3 OFDMA modulation frame structure as per 16e standard

The Frame structures mentioned here only for demonstrating the concept and it differs in the actual wimax system. Both OFDM and OFDMA is used to achieve high data rate transmission over the air. With OFDMA system can support more subscribers with sub channelization concept compare to OFDM. Both OFDM and OFDMA is implemented using IFFT and FFT operation at transmitter and receiver respectively. For OFDM entire input of IFFT is occupied fully by either subscriber staion or Base Station. For OFDMA part of input values (consecutively) is occupied by Subscriber station and at rest of the input positions zeros or nulls are inserted. Same is done with other subscribers and so on.

2.3.3 GFDM:

Generalized Frequency Division Multiplexing (GFDM) is one of the nonorthogonal multi-carrier transmission methods that has been considered for 5G systems. It provides low OOB radiation and frequency localization due to variable pulse shaping filters, making it an attractive choice for IoT and cognitive radios operating in TV white spaces. Studies have shown its superiority over OFDM due to low OOB radiation and low PAPR. As stated above, GFDM employs variable pulse shaping filters to achieve frequency localization. This localization allows the waveform to fit into narrow spectral holes eliminating interference to adjacent frequency bands. An ideal pulse shape needs to attenuate very sharply both in frequency and time domains to avoid overlap with adjacent carriers/symbols to avoid intercarrier interference (ICI) and ISI. However, such pulse shapes do not exist and thus compromise has to be made with attenuation depending on the channel characteristics. Such filters also affect orthogonality between the subcarriers resulting in ICI/ISI, which can be addressed by efficient detection techniques at the receiver side. The GFDM transceiver is similar to an OFDM transceiver except it uses pulse-shaped filters for each subcarrier and a tail biting* technique. A few Rx-filter (receive filter) types come in handy for GFDM, that is, matched filter receiver, zero forcing receiver, and minimum mean square error (MMSE) receiver with varying performances.

2.3.4 NOMA:

Today's LTE and 4G (LTE-Advanced) technologies are based on an OMA scheme, that is, OFDM. It is widely known that OFDM suffers from high PAPR, can introduce ICI due to loss in subcarrier orthogonality, and has some other impairments as well. Thus, to improve spectral efficiency, a nonorthogonal scheme, namely NOMA (Non-Orthogonal Multiple Access), has been considered for 5G. NOMA brings an additional attribute to the picture, that is, power which has not been considered to differentiate users by any currently deployed multiple access scheme. In NOMA, multiple users can transmit at the same time using the same code and frequency but with different power levels. In this access method, multiple users are multiplexed in the power domain on the transmitting end and on the receiving side, SIC (Successive Interference Cancellation) can be used for multi-user signal separation. This power sharing reduces the amount of power allocated to each user, therefore, users with high channel gains are assigned less power as compared to users with lower channel gains. The performance gain compared to OMA increases when the difference in channel gains (e.g., path loss between user terminals) is large. NOMA superposes multiple users in the power domain (forming a superposition coding) while enabling user separation at the receiving end through SIC. NOMA introduces additional complexity and delay due to the use of SIC and the performance gain is also insignificant at low SNR. NOMA is suitable for both eMBB and mMTC (Massive Machine-Type Communications) types of services, but perhaps not for URLLC due to the inherent delay associated with SIC.

Non-orthogonal multiple access (NOMA) has been recently recognized as a promising multiple access technique to significantly improve the spectral efficiency of mobile communication networks. In 1G, 2G, and 3G, frequency division multiple access (FDMA), time division multiple access and code division multiple access were introduced, respectively. In Long-Term Evolution (LTE) and LTE-Advanced, orthogonal frequency division multiple access (OFDMA) and single-carrier (SC)-FDMA are adopted as an orthogonal multiple access(OMA) approach. Such an orthogonal design has the benefit that there is no mutual interference among users, and therefore good system-level performance can be achieved even with simplified receivers. However, none of these techniques can meet the high demands of future radio access systems such as 5G. The increasing demand of mobile Internet and the Internet of Things poses challenging requirements for 5G wireless communications, such as high spectral efficiency and massive connectivity. In this article, a promising technology, non-orthogonal multiple access (NOMA), is discussed, which can address some of these challenges for 5G.

2.3.4.1 Basic Concept:

Figure below illustrates downlink NOMA for the case of one BS (base station) and two UE (user equipment). In downlink NOMA, the transmit signal from the BS and the received signal at both UE receivers is composed of a superposition of the transmit signals of both UEs. Thus multi-user signal separation needs to be implemented at the UE side so that each UE can retrieve its signal and decode its own data. This can be achieved by non-linear receivers such as maximum likelihood detection or SIC (Successive Interference Cancellation).

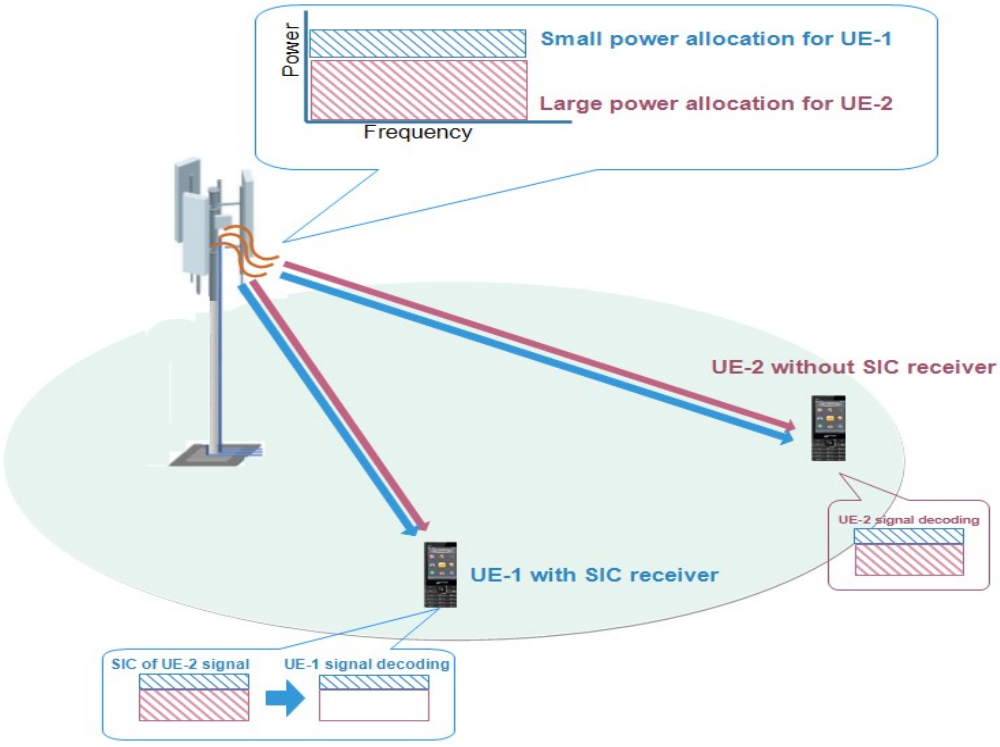


Figure 2.4: Downlink NOMA

For the case of SIC, the optimal order for decoding is in the order of the decreasing channel gain normalized by noise and ICI power. Based on this order, we can actually assume that any user can correctly decode the signals of other users whose decoding order comes before the corresponding user. In a two-UE case, assuming that, UE-2 does not perform interference cancellation since it comes first in the decoding order. UE-1 first decode UE-2 signal, and subtracts its component from total received signal, and thus it gets its own signal component and decodes it, without interference from UE-2 signal. NOMA uses the power domain to separate signals from each other. NOMA gives a new dimension in which signals can be separated and given access to a base station. This technique that has not been used within 2G, 3G or 4G before.

2.3.4.2 Comparison with OFDMA:

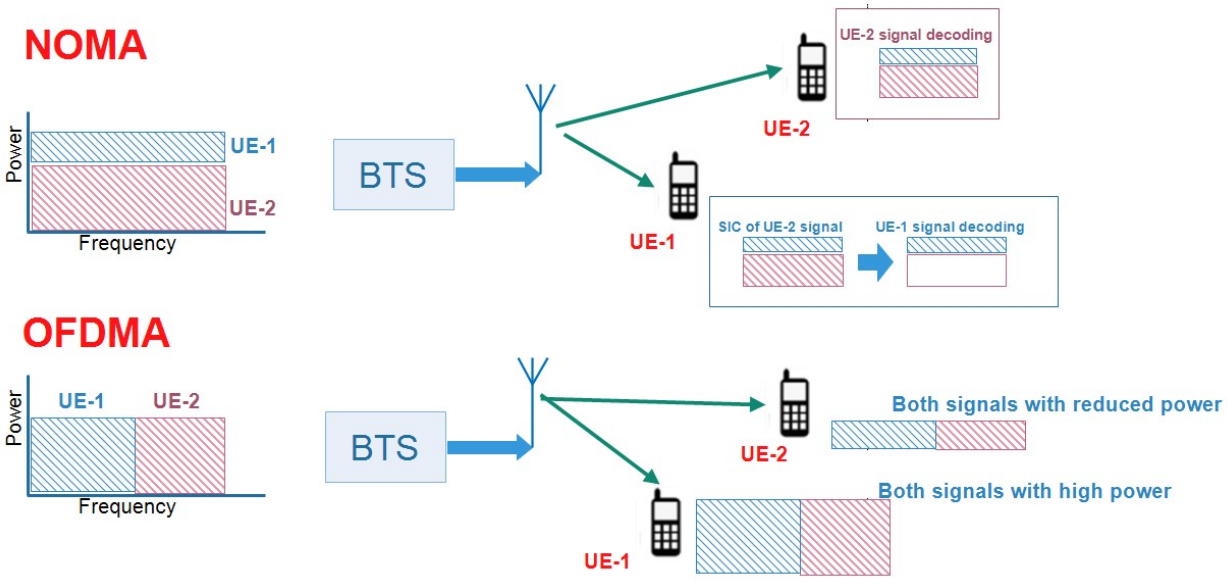


Figure 2.5: Comparison with OFDMA

In OFDMA, different UE signals are transmitter at different frequency resources, but in case of NOMA, different UE signals are transmitted at same frequency but at different power levels depending upon the position of UE in the cell. The performance gain of NOMA compared to that of OFDMA increases when the difference in channel gain the path loss between UEs is large. According to this simple two-UE case, NOMA provides a higher sum rate than OFDMA. In fact, the cell-center UE gains in terms of rate since this UE is bandwidth-limited and thus benefits more from being able to use double bandwidth, even if this comes at the price of much lower transmit power. Meanwhile, the cell-edge UE also gains in terms of rate since it is power-limited; its transmit power is only slightly reduced under NOMA but its transmit bandwidth can be doubled.

2.3.4.3 Applications:

As a promising 5G technique, NOMA has been shown to be compatible with other key enabling techniques for 5G communications. For example, the heterogeneous network architecture will play an important role in 5G networks, where macro base stations and small cell base stations cooperate for spectrum sharing. The NOMA is beneficial for heterogeneous networks, as more users can be served in a small cell by exploiting the NOMA principle. At the same time, the applications of NOMA to machine-to-machine (M2M) communications, ultra-dense networks (UDN), and massive machine type communications (mMTC) are being studied, respectively, where the use of NOMA can effectively support massive connectivity and the IoT functionality of 5G. With its distinct features, NOMA stays as the strongest candidate for the future 5G networks. There are, however, still some challenges for successful implementation of NOMA. First of all, it requires high computational power to run SIC algorithms particularly for high number of users at high data rates. Second, power allocation optimization remains as a challenging problem, particularly when the UEs are moving fast in the network. Finally, SIC receiver is sensitive to cancellation errors which can easily occur in fading channels. It can be implemented with some other diversity techniques like multiple-input-multiple-output (MIMO) or with coding schemes in order to increase the reliability and accordingly reduce the decoding errors. The current state of the art for NOMA, however, is still far from its potential and requires further investigation.

2.3.5 UFMC

The UFMC or Universal Filtered Multicarrier is a modification of the well-known waveform CP-OFDM. The term UF-OFDM (Universal Filtered OFDM) is also used synonymously with UFMC. In CP-OFDM, symbols are separated using CP and the entire frequency band is digitally filtered as a whole. UFMC, however, applies filtering on a per sub-band (i.e., a block of subcarriers) basis and avoids use of CP. The sub-band wise filtering approach was investigated since time-frequency misalignments normally occur between blocks of subcarriers (for example, sub-band wise resource allocation of different uplink users). Additionally, as the filters are broader in frequency, these become shorter in time, providing better communications in short bursts which is required for mMTC/IoT applications. It may be noted that the use of zero padding instead of CP improves spectral efficiency; however, it makes UFMC more sensitive to time misalignment as compared to the CP-OFDM waveform.

2.3.6 FBMC:

Filter bank multicarrier (FBMC) is one of the potential 5G waveforms where filtering is considered at a very granular level, that is, on a per subcarrier basis. In simple terms, FBMC represents a multi-carrier system where single subcarrier signals are individually filtered with prototype filters [30]. FBMC has been proposed for cognitive radio applications.

For a typical multi-access system to work, the receiver (FFT) must be perfectly aligned in time with the transmitter (IFFT). During multipath propagation, the multicarrier symbols overlap at the receiver input resulting in ISI. The ISI further results in the loss of orthogonality of the carriers making demodulation harder with just the FFT. FBMC addresses this challenge by

adding some additional processing to the FFT while keeping the timing and the symbol duration as it is. This additional processing together with the FFT constitutes a bank of filters. The FBMC approach is different from both OFDM where filtering is applied on the entire frequency band and UFMC which filters on a sub-band basis. Thus, instead of having sinc-pulses like OFDM, the subcarriers have an appropriate shape according to the filter design and with negligible sidelobes.

However, the prototype filters are very narrow in frequency, necessitating rather long filter lengths (typically 3–4 times the basic multicarrier symbol length). The longer filter lengths require long ramp up and ramp down areas to address bursty data transmissions. The subcarrier filtering eases spectrum sharing and spectrum sensing, making FBMC highly applicable for cognitive radio networks. FBMC also offers higher robustness against Doppler and time and frequency impairments compared to OFDM due to the use of appropriate filters. It also provides higher spectral efficiency as it does not use a cyclic prefix.

2.3.7 SCMA:

New Radio (NR) is a newly approved study item in 3GPP, focusing on the design of the next generation (5G) air interface. 5G air interface is intended for having faster access, higher transmission rates, support of larger user density, and overall better user experience. At the same time, 5G connects to new vertical industries and devices, resulting in creating new application scenarios like mMTC and URLLC services. This happens by supporting a huge number of devices and enabling mission-critical transmissions through super high reliability and super low latency requirement, respectively.

SCMA is a non-orthogonal multiple-access technique that is being used for possible use with 5G and other developed communications systems. The target is that SCMA, i.e. Sparse Code Multiple Access, will upgrade spectral efficiency of wireless radio access. In many respects, SCMA can be considered as a combination of CDMA, Code Division Multiple Access and OFDMA, Orthogonal Frequency Division Multiple Access.

SCMA can be considered as a code division multiple access scheme, which is described by sparse codebooks. The codebooks are built based on multidimensional constellations, and the shaping gain helps it outperform the traditional spread code based schemes. In SCMA, multiple users will transmit on the same resource blocks with different codebooks. With sparse codebooks, the collision between users is reduced, thus SCMA is resilient to inter-user interference. The sparsity is also benefit for the receiver complexity, and the message passing algorithm can be applied to achieve near-optimal performance.

2.3.7.1 SCMA Codebook Mapping:

SCMA can be of similar layer mapping as LTE, which means that one or more SCMA layers can be assigned to a user/data stream. The difference is that at each SCMA layer, the SCMA will also do mapping from information bits to codewords, in other words, the SCMA modulator maps input bits to a more complicated multi-dimensional codeword chosen from a

layer-specific SCMA codebook. SCMA codewords are sparse, which means that only a few of their entries are non-zero and the rest of them are zero. All SCMA codewords corresponding to a SCMA layer have a unique location of non-zero entries, referred to as sparsity pattern for simplicity.

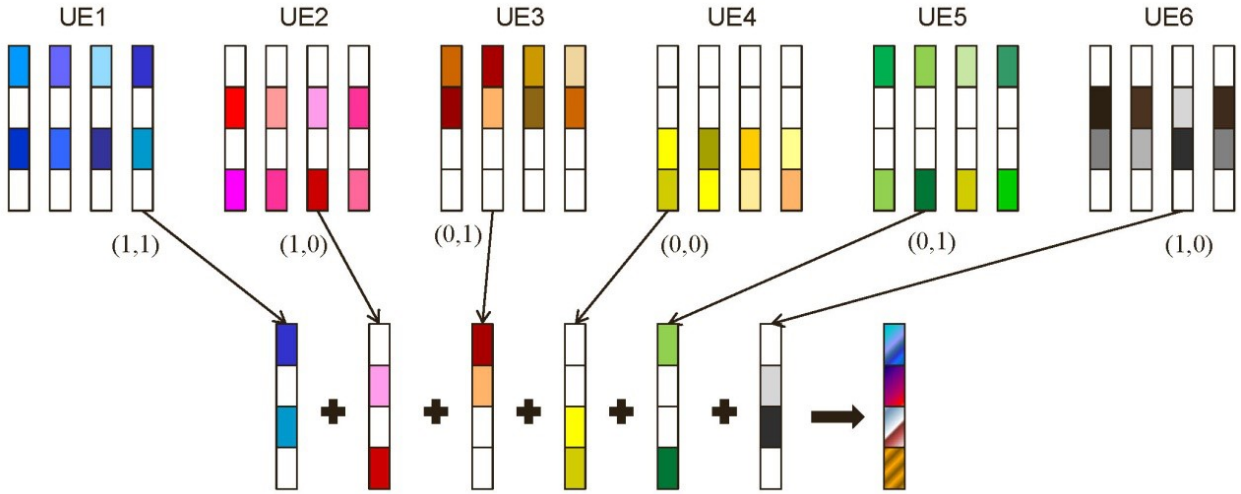


Figure 2.6: SCMA Codebook Mapping

Example of a CodeBook: As per the below figure, a codebook set containing 6 codebooks for transmitting 6 data layers, each of the codebook has 8 multi-dimensional complex codewords that correspond to 8 points of constellation, respectively. The length of every codeword is four (4), which is exactly the same as the spreading length. Upon transmission, the codeword of each layer is selected based on the input bit sequence. In the downlink, as shown in the below figure, the codewords from different layers are combined before the OFDM modulator. In the uplink, for a single layer UE transmission, each SCMA codeword is first fed into the OFDM modulator resulting in multiple independent SCMA layers over the air transmissions from different users.

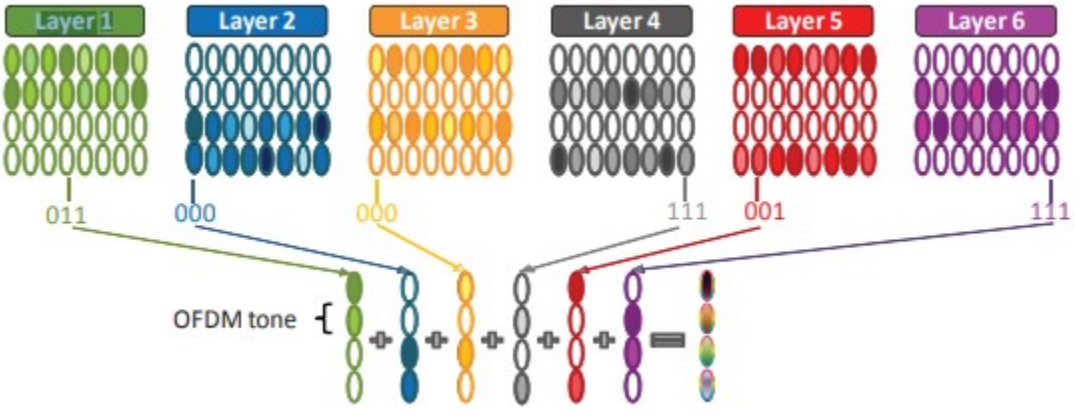


Figure 2.7: SCMA codebook bit-to-codeword mapping

Figure 2.8: SCMA 8-point codebook

SCMA Codebook Design:

The design of SCMA codebook is founded on combined optimization of the sparse spreading pattern design and the multidimensional modulation design.

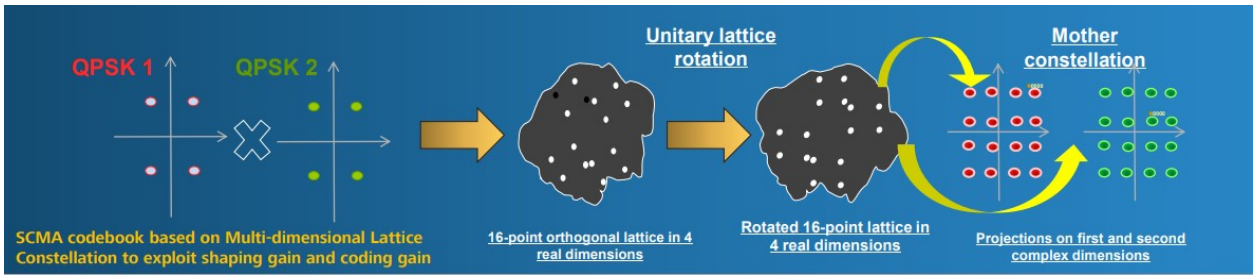


Figure 2.9: SCMA Codebook Design

2.3.7.2 Multiple Access with SCMA

An example of multiple access of 6 users with the SCMA layer-specific codebooks. One SCMA codebook is given to each user (in the given example here, user i takes codebook for layer i, i = 1, 2, …, 6). After FEC encoder (e.g., LDPC encoder), each user's coded bits are then mapped to the SCMA codeword according to its assigned codebook. The SCMA codewords are further combined over OFDM tones and symbols are transmitted in terms of SCMA blocks, similar to resource block concept in LTE.

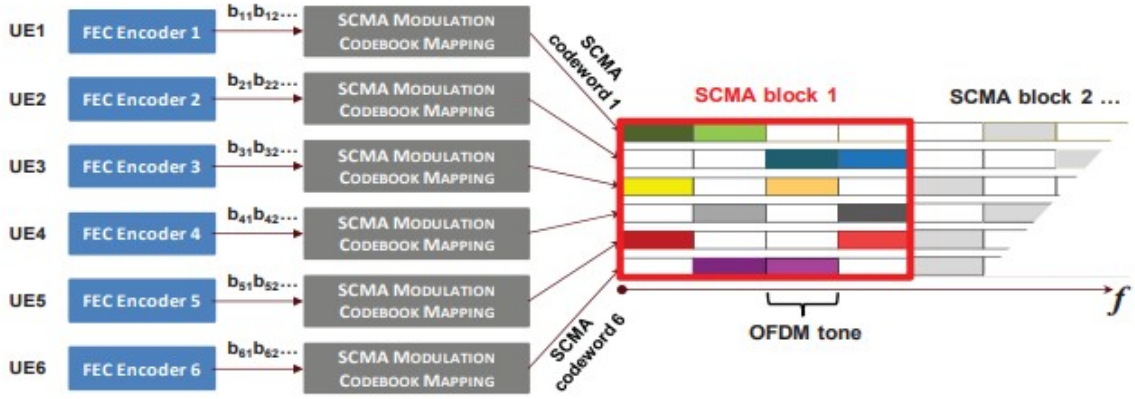


Figure 3.10: Multiple Access with SCMA

The Main Characteristics of Multiple Access with SCMA Code Domain Non-orthogonal Signal Superposition: It creates a superposition of multiple symbols from different users on each Resource Element (RE). For instance, in the given figure, on RE 1, symbols from UE1, 3, and 5 are overlapped with each other. The super-position pattern on each RE can be statically or semi-statically configured. Sparse Spreading: SCMA uses sparse spreading to decrease the number of symbol collisions. For example, in the above figure, there are 3 symbols from different UEs are colliding over each RE, instead of 6 (in the case of non-sparse spreading).

2.3.7.3 Applications:

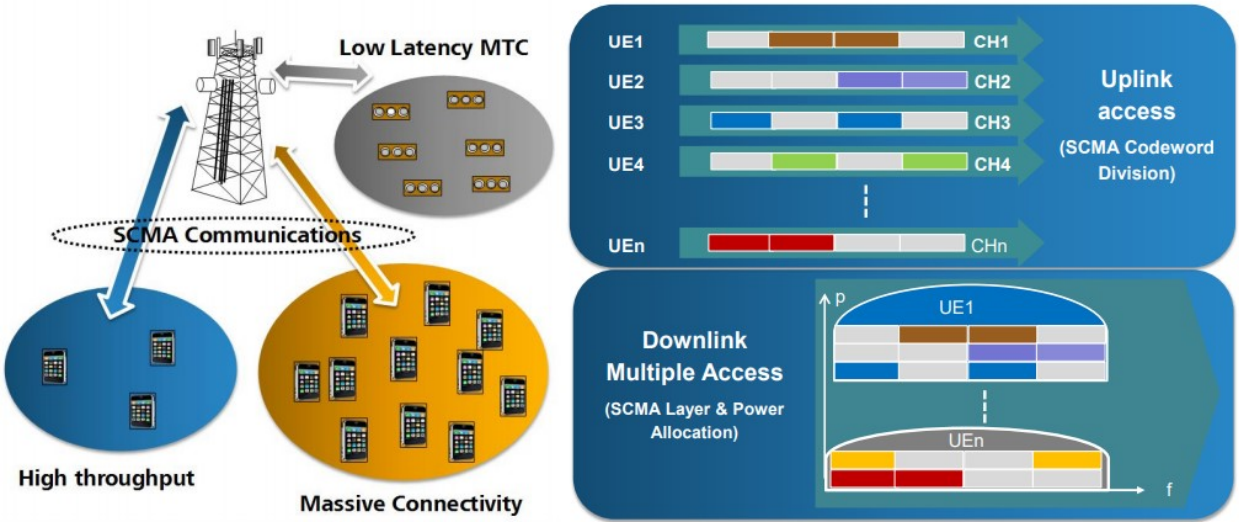


Figure 3.11: SCMA Application Scenarios

Since the access to resources with orthogonal multiple access technology is proportionate to the number of users, it can't meet the 5G capacity, low latency access requirement, and massive connectivity. So, non-orthogonal multiple access will be the 5G Multiple Access research focuses access. SCMA is designed to generate demand should 5G a non-orthogonal multiple access technology.

Reference Books:

1. Saad Z. Asif, "5G Mobile Communications Concepts and Technologies, CRC Press, 1st Edition, 2019.
2. Erik Dahlman, Stefan Parkvall, Johan Skold "5G NR: The Next Generation Wireless Access Technology", Academic Press, 1st Edition, 2018.
3. Jonathan Rodriguez, "Fundamentals 5G Mobile Networks", John Wiley & Sons, 1st Edition, 2015.
4. Long Zhao, Hui Zhao, Kan Zheng, Wei Xiang, "Massive MIMO in 5G Networks: Selected Applications", Springer, 1st Edition, 2018.
5. Robert W. Heath Jr., Angel Lozano, "Foundations of MIMO Communication", Cambridge University Press, 1st Edition, 2019.
6. R. Vannithamby and S. Talwar, "Towards 5G: Applications, Requirements and Candidate Technologies", John Willey & Sons, 1st Edition, 2017.

Questions to Practice:

PART -A

1	Demonstrate OFDMA with frame structure
2	Interpret the use of PAPR where Generalized Frequency Division Multiplexing leads to multicarrier filters.
3	Demonstrate OFDM with frame structure
4	Demonstrate GFDM with frame structure
5	Explain the Multiple access technique are built based on multidimensional constellations, and the shaping gain helps it outperform the traditional spread code based schemes

PART-B

1	Identify how NOMA can overcome the above statement.
2	. Examine Multiple Access Techniques that are used in 5g Technology.
3	The existing orthogonal frequency division multiple access (OFDMA) technology cannot support massive connectivity, the severe out-of-band (OOB) power leakage not only waste spectrum resources but also hinders the application of OFDM in the fragmented spectrum. Relate how FBMC and SCMA can overcome the scenario.
4	Contrast OFDM and OFDMA in preventing signal interference and helping you to achieve the most streamlined network experience possible.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF ELECTRICAL AND ELECTRONICS

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

UNIT – III – RADIO ACCESS NETWORK FOR 5G – SECA3020

UNIT-III (RADIO ACCESS NETWORK FOR 5G NR)

5G NR requirements - 5G Core Network Architecture - Radio-Access Network (RAN)- Radio Protocol Architecture -User Plane Protocols-Radio Link Control - Medium-Access Control – Physical Layer functions -Control Plane Protocols - Network Slicing- RAN virtualization-Spectrum Management in 5G

3.1 Introduction :

The timeline for the NR development within 3GPP is shown in figure 3.1. The technical work on NR was initiated in the spring of 2016 as a study item in 3GPP release 14, based on a kick-off workshop in the fall of 2015. During the study item phase, different technical solutions were studied, but given the tight time schedule, some technical decisions were taken already in this phase. The work continued into a work item phase in release 15, resulting in the first version of the NR specifications available by the end of 2017, before the closure of 3GPP release 15 in mid- 2018. The reason for the intermediate release of the specifications, before the end of release-15, is to meet commercial requirements on early 5G deployments. The first specification from December 2017, which is the focus of this book, is limited to non-standalone NR operation (see Chapter 6), implying that NR devices rely on LTE for initial access and mobility. The final release-15 specifications support standalone NR operation as well. The difference between standalone and non-standalone primarily affects higher layers and the interface to the core network; the basic radio technology is the same in both cases. During the development of release 15, the focus was on eMBB and (to some extent) URLLC type of services. For massive machine-type communication (mMTC), LTE-based technologies such as eMTC and NB-IoT can be used with excellent results. The support for LTE-based massive MTC on a carrier overlapping with an NR carrier has been accounted for in the design of NR, resulting in an integrated overall system capable of handling a very wide range of services. Native NR support for extended mMTC, as well as special technology features such as direct device-to-device connectivity, in 3GPP referred to as sidelink transmission, will be addressed in later releases.

In parallel to the work on the NR radio-access technology in 3GPP, a new 5G core network has been developed, responsible for functions not related to the radio access but needed for providing a complete network. However, it is possible to connect the NR radio-access network also to the legacy LTE core network known as the Evolved Packet Core (EPC). In fact, this is the case when operating NR in non-standalone mode where LTE and EPC handle functionality like connection set-up and paging and NR primarily provides a data-rate and capacity booster. Later releases will introduce standalone operation with NR connecting to the 5G core. The remaining part of this chapter provides an overview of NR radio access including basic design principles and the most important technology components

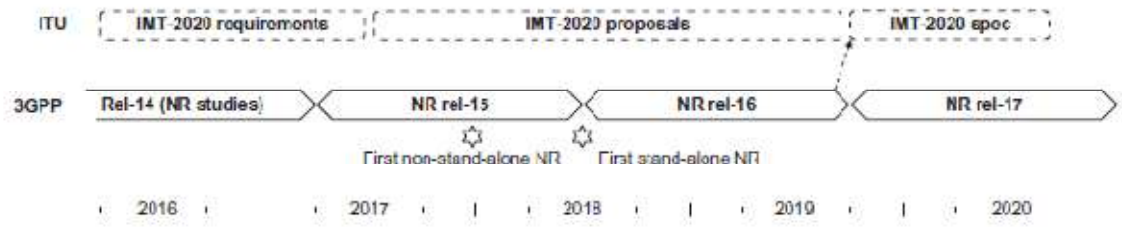


Figure 3.1 3GPP Timeline

of NR release 15. The chapter can either be read on its own to get a high-level overview of NR, or as an introduction to the subsequent, which provide a detailed description of the NR.

Compared to LTE, NR provides many benefits. Some of the main ones are:

- exploitation of much higher-frequency bands as a mean to obtain additional spectra to support very wide transmission bandwidths and the associated high data rates;
- ultra-lean design to enhance network energy performance and reduce interference;
- forward compatibility to prepare for future, yet unknown, use cases and technologies;
- low latency to improve performance and enable new use cases; and
- a beam-centric design enabling extensive usage of beamforming and a massive number of antenna elements not only for data transmission (which to some extent is possible in LTE) but also for control-plane procedures such as initial access. The first three can be classified as design principles (or requirements on the design) and will be discussed first, followed by a discussion of the key technology components applied to NR.

3.1.1 HIGHER-FREQUENCY OPERATION AND SPECTRUM FLEXIBILITY

One key feature of NR is a substantial expansion in terms of the range of spectra in which the radio-access technology can be deployed. Unlike LTE, where support for licensed spectra at 3.5 GHz and unlicensed spectra at 5 GHz are just being introduced, NR supports licensed-spectrum operation from below 1 GHz up to 52.6 GHz already from its first release, with extension to unlicensed spectra also already being planned for.

Operation at mm-wave frequencies offers the possibility for large amounts of spectrum and associated very wide transmission bandwidths, thereby enabling very high traffic capacity and extreme data rates. However, higher frequencies are also associated with higher radio-channel attenuation, limiting the network coverage. Although this can partly be compensated for by means of advanced multiantenna transmission/reception, which is one of the motivating factors for the beam-centric design in NR, a substantial coverage disadvantage remains, especially in non-line-of-sight and outdoor-to-indoor propagation conditions. Thus, operation in lower-frequency bands will remain a vital component for wireless communication also in the 5G era. Especially, joint operation in lower and higher spectra, for example 2 GHz and 28 GHz, can provide substantial benefits. A higher-frequency layer, with access to a large amount of spectra can provide service to a large fraction of the users despite the more limited coverage. This will reduce the load on the more bandwidth-constrained lower-frequency spectrum, allowing the use of this to focus on the worst-case users [66].

Another challenge with operation in higher-frequency bands is the regulatory aspects. For non-technical reasons, the rules defining the allowed radiation changes at 6 GHz, from a SAR-based limitation to a more EIRP-like limitation. Depending on the device type (handheld, fixed, etc.), this may result in a reduced transmission power, making the link budget more challenging than what propagation conditions alone may indicate and further stressing the benefit of combined low-frequency/high-frequency operation.

3.1.2 ULTRA-LEAN DESIGN

An issue with current mobile-communication technologies is the amount of transmissions carried by network nodes regardless of the amount of user traffic. Such signals, sometimes referred to as "always-on" signals, include, for example, signals for base-station detection, broadcast of system information, and always-on reference signals for channel estimation. Under the typical traffic conditions for which LTE was designed, such transmissions constitute only a minor part of the overall network transmissions and thus have a relatively small impact on the network performance. However, in very dense networks deployed for high peak data rates, the average traffic load per network node can be expected to be relatively low, making the always-on transmissions a more substantial part of the overall network transmissions.

The always-on transmissions have two negative impacts:

- they impose an upper limit on the achievable network energy performance; and
- they cause interference to other cells, thereby reducing the achievable data rates.

The ultra-lean design principle aims at minimizing the always-on transmissions, thereby enabling higher network energy performance and higher achievable data rates. In comparison, the LTE design is heavily based on cell-specific reference signals, signals that a device can assume are always present and use for channel estimation, tracking, mobility measurements, etc. In NR, many of these procedures have been revisited and modified to account for the ultra-lean design principle. For example, the cell-search procedures have been redesigned in NR compared to LTE to support the ultra-lean paradigm. Another example is the demodulation reference-signal structure where NR relies heavily on reference signals being present only when data are transmitted but not otherwise.

3.1.3 FORWARD COMPATIBILITY

An important aim in the development of the NR specification was to ensure a high degree of forward compatibility in the radio-interface design. In this context, forward compatibility implies a radio-interface design that allows for substantial future evolution, in terms of introducing new technology and enabling new services with yet unknown requirements and characteristics, while still supporting legacy devices on the same carrier. Forward compatibility is inherently difficult to guarantee. However, based on experience from the evolution of previous generations, 3GPP agreed on some basic design principles related to NR forward compatibility as quoted from

- Maximizing the amount of time and frequency resources that can be flexibly utilized or that can be left blank without causing backward compatibility issues in the future;

- Minimizing transmission of always-on signals;

• Confining signals and channels for physical layer functionalities within a configurable/allocable time/frequency resource.

According to the third bullet one should, as much as possible, avoid having transmissions on time/frequency resources fixed by the specification. In this way one retains flexibility for the future, allowing for later introduction of new types of transmissions with limited constraints from legacy signals and channels. This differs from the approach taken in LTE where, for example, a synchronous hybrid-ARQ protocol is used, implying that a retransmission in the uplink occurs at a fixed point in time after the initial transmission. The control channels are also vastly more flexible in NR compared to LTE in order not to unnecessarily block resources.

Note that these design principles partly coincide with the aim of ultra-lean design as described above. There is also a possibility in NR to configure reserved resources, that is, time-frequency resources that, when configured, are not used for transmission and thus available for future radio-interface extensions. The same mechanism can also be used for LTE-NR coexistence in the case of overlapping LTE and NR carriers.

3.1.4 TRANSMISSION SCHEME, BANDWIDTH PARTS, AND FRAME STRUCTURE

Similar to LTE, OFDM was found to be a suitable waveform for NR due to its robustness to time dispersion and ease of exploiting both the time and frequency domain when defining the structure for different channels and signals. However, unlike LTE where DFT-precoded OFDM is the sole transmission scheme in the uplink, NR uses conventional, that is, non-DFT-precoded OFDM, as the baseline uplink transmission scheme due to the simpler receiver structures in combination with spatial multiplexing and an overall desire to have the same transmission scheme in both uplink and downlink. Nevertheless, DFT-precoding can be used as a complement in the uplink for similar reasons as in LTE, namely to enable high power-amplifier efficiency on the device side by reducing the cubic metric. Cubic metric is a measure of the amount of additional power back-off needed for a certain signal waveform. To support a wide range of deployment scenarios, from large cells with sub- 1 GHz carrier frequency up to mm-wave deployments with very wide spectrum allocations, NR supports a flexible OFDM numerology with subcarrier spacings ranging from 15 kHz up to 240 kHz with a proportional change in cyclic prefix duration. A small subcarrier spacing has the benefit of providing a relatively long cyclic prefix in absolute time at a reasonable overhead while higher subcarrier spacings are needed to handle, for example, the increased phase noise at higher carrier frequencies. Up to 3300 subcarriers are used although the maximum total bandwidth is limited to 400 MHz, resulting in the maximum carrier bandwidths of 50/100/200/400 MHz for subcarrier spacings of 15/30/60/120 kHz, respectively. If even larger bandwidths are to be supported, carrier aggregation can be used.

Although the NR physical-layer specification is band-agnostic, not all supported numerologies are relevant for all frequency bands. For each frequency band, radio requirements are therefore defined for a subset of the supported numerologies. The frequency range 0.456 GHz is commonly referred to as frequency range 1 (FR1) in the specifications, while the range 24.2552.6 GHz is known as FR2. Currently, there is no NR spectrum identified between 6 GHz

and 24.25 GHz. However, the basic NR radio-access technology is spectrum agnostic and the NR specifications can easily be extended to cover additional spectra, for example, spectra from 6 GHz up to 24.25 GHz. In LTE, all devices support the maximum carrier bandwidth of 20 MHz. However, given the very wide bandwidths possible in NR, it is not reasonable to require all devices to support the maximum carrier bandwidth. This has implications on several areas and requires a design different from LTE, for example the design of control channels as discussed later. Furthermore, NR allows for device-side receiver-bandwidth adaptation as a means to reduce the device energy consumption. Bandwidth adaptation refers to the use of a relatively modest bandwidth for monitoring control channels and receiving medium data rates, and dynamically opens up a wideband receiver only when needed to support very high data rates. To handle these two aspects NR defines bandwidth parts that indicate the bandwidth over which a device is currently assumed to receive transmissions of a certain numerology. If a device is capable of simultaneous reception of multiple bandwidths parts, it is in principle possible to, on a single carrier, mix transmissions of different numerologies for a single device, although release 15 only supports a single active bandwidth part at a time.

3.1.5 DUPLEX SCHEMES

The duplex scheme to use is typically given by the spectrum allocation at hand. For lower-frequency bands, allocations are often paired, implying frequencydivision duplex (FDD) as illustrated in Fig. 5.4. At higher-frequency bands, unpaired spectrum allocations are increasingly common, calling for time-division duplex (TDD). Given the significantly higher carrier frequencies supported by NR compared to LTE, efficient support for unpaired spectra is an even more critical component of NR, compared to LTE. NR can operate in both paired and unpaired spectra using one common frame structure, unlike LTE were two different frame structures were used (and later expanded to three when support for unlicensed spectra was introduced in release 13). The basic NR frame structure is designed such that it can support both halfduplex and full-duplex operation. In half duplex, the device cannot transmit and receive at the same time. Examples hereof are TDD and half-duplex FDD. In full-duplex operation, on the other hand, simultaneous transmission and reception is possible with FDD as a typical example.

3.1.6 LOW-LATENCY SUPPORT

The possibility for very low latency is an important characteristic of NR and has impacted many of the NR design details. One example is the use of “frontloaded” reference signals and control signaling. By locating the reference signals and downlink control signaling carrying scheduling information at the beginning of the transmission and not using time-domain interleaving across OFDM symbols, a device can start processing the received data immediately without prior buffering, thereby minimizing the decoding delay. The possibility for transmission over a fraction of a slot, sometimes referred to as “mini-slot” transmission, is another example. The requirements on the device (and network) processing times are tightened significantly in NR compared to LTE. As an example, a device has to respond with a hybrid-ARQ acknowledgment of approximately one slot (or even less depending on device capabilities) after receiving a downlink data transmission. Similarly, the time from grant reception to uplink data transfer is in the same range.

The higher-layer protocols MAC and RLC have also been designed with low latency in mind with header structures chosen to enable processing without knowing the amount of data to transmit. This is especially important in the uplink direction as the device may only have a few OFDM symbols after receiving the uplink grant until the transmission should take place. In contrast, the LTE protocol design requires the MAC and RLC protocol layers to know the amount of data to transmit before any processing can take place, which makes support for a very low latency more challenging.

3.1.7 SCHEDULING AND DATA TRANSMISSION

One key characteristic of mobile radio communication is the large and typically rapid variations in the instantaneous channel conditions stemming from frequency-selective fading, distance-dependent path loss, and random interference variations due to transmissions in other cells and by other devices. Instead of trying to combat these variations, they can be exploited through channel-dependent scheduling where the time-frequency resources are dynamically shared between users. Dynamic scheduling is used in LTE as well and on a high level, the NR scheduling framework is similar to the one in LTE. The scheduler, residing in the base station, takes scheduling decisions based on channel-quality reports obtained from the devices. It also takes different traffic priorities and quality-of-service requirements into account when forming the scheduling decisions sent to the scheduled devices. Each device monitors several physical downlink control channels (PDCCHs), typically once per slot, although it is possible to configure more frequent monitoring to support traffic requiring very low latency. Upon detection of a valid PDCCH, the device follows the scheduling decision and receives (or transmits) one unit of data known as a transport block in NR. In the case of downlink data transmission, the device attempts to decode the downlink transmission. Given the very high data rates supported by NR, channel coding data transmission is based on low-density parity-check (LDPC) codes. LDPC codes are attractive from an implementation perspective, especially at higher code rates where they can offer a lower complexity than Turbo codes as used in LTE.

Hybrid automatic repeat-request (ARQ) retransmission using incremental redundancy is used where the device reports the outcome of the decoding operation to the base station (see Chapter 13 for details). In the case of erroneously received data, the network can retransmit the data and the device combines the soft information from multiple transmission attempts. However, retransmitting the whole transport block could in this case become inefficient. NR therefore supports retransmissions on a finer granularity known as code-block group (CBG). This can also be useful when handling preemption. An urgent transmission to a second device may use only one or a few OFDM symbols and therefore cause high interference to the first device in some OFDM symbols only. In this case it may be sufficient to retransmit the interfered CBGs only and not the whole data block. Handling of preempted transmission can be further assisted by the possibility to indicate to the first device the impacted time-frequency resources such that it can take this information into account in the reception process. Although dynamic scheduling is the basic operation of NR, operation without a dynamic grant can be configured. In this case, the device is configured in advance with resources that can be used for uplink data transmission (or downlink data reception). Once a device has data available it can immediately commence uplink

transmission without going through the scheduling requestgrant cycle, thereby enabling lower latency.

3.2 RADIO-INTERFACE ARCHITECTURE

In parallel to the work on the NR (New Radio) radio-access technology in 3GPP, the overall system architectures of both the Radio-Access Network (RAN) and the Core Network (CN) were revisited, including the split of functionality between the two networks.

The RAN is responsible for all radio-related functionality of the overall network including, for example, scheduling, radio-resource handling, retransmission protocols, coding, and various multi-antenna schemes.

The 5G core network is responsible for functions not related to the radio access but needed for providing a complete network. This includes, for example, authentication, charging functionality, and setup of end-to-end connections. Handling these functions separately, instead of integrating them into the RAN, is beneficial as it allows for several radio-access technologies to be served by the same core network. However, it is possible to connect the NR radio-access network also to the legacy LTE (Long-Term Evolution) core network known as the Evolved Packet Core (EPC). In fact, this is the case when operating NR in non-standalone mode, where LTE and EPC handle functionality like connection set-up and paging. Later releases will introduce standalone operation with NR connecting to the 5G core, as well as LTE connecting to the 5G core. Thus, the LTE and NR radio-access schemes and their corresponding core networks are closely related, unlike the transition from 3G to 4G where the 4G LTE radio-access technology cannot connect to a 3G core network.

Although this book focuses on the NR radio access, a brief overview of the 5G core network, as well as how it connects to the RAN, is useful as a background.

3.2.1 5G CORE NETWORK

The 5G core network builds upon the EPC with three new areas of enhancement compared to EPC: service-based architecture, support for network slicing, and control-plane/user-plane split.

A service-based architecture is the basis for the 5G core. This means that the specification focuses on the services and functionalities provided by the core network, rather than nodes as such. This is natural as the core network today is already often highly virtualized with the core network functionality running on generic computer hardware. Network slicing is a term commonly seen in the context of 5G. A network slice is a logical network serving a certain business or customer need and consists of the necessary functions from the service-based architecture configured together. For example, one network slice can be set up to support mobile broadband applications with full mobility support, similar to what is provided by LTE, and another slice can be set up to support a specific non-mobile, latency-critical industry-automation application.

These slices will all run on the same underlying physical core and radio networks, but, from the end-user application perspective, they appear as independent networks. In many aspects it is similar to configuring multiple virtual computers on the same physical computer. Edge computing, where parts of the end-user application run close to the core network edge to provide low latency, can also be part of such a network slice. Control-plane/user-plane split is emphasized in the 5G core network architecture, including independent scaling of the capacity of the two. For example, if more control plane capacity is need, it should be straightforward to add it without affecting the user-plane of the network.

On a high level, the 5G core uses a service-based representation, where the services and functionalities are High-level core network architecture (service-based description).

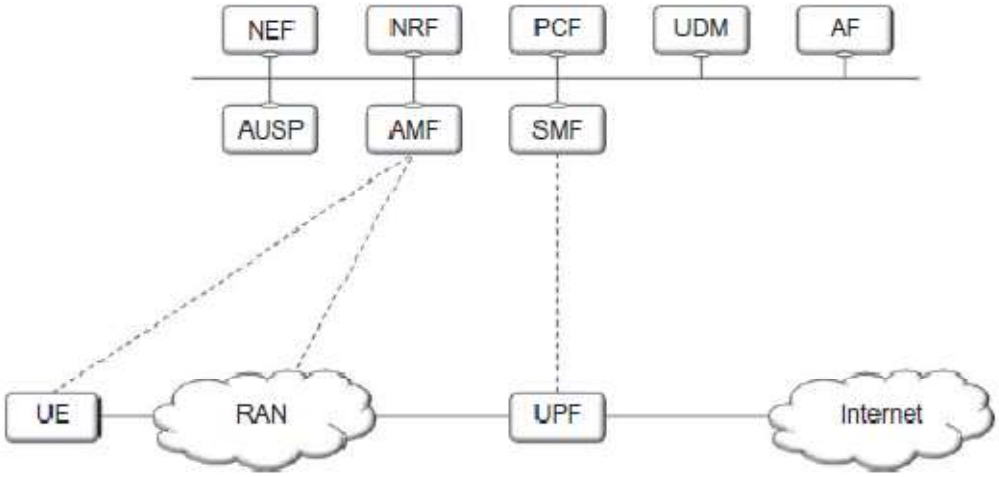


Figure 3.2: High-level core network architecture (service-based description)

The user-plane function consists of the User Plane Function (UPF) which is a gateway between the RAN and external networks such as the Internet. Its responsibilities include packet routing and forwarding, packet inspection, quality-of service handling and packet filtering, and traffic measurements. It also serves as an anchor point for (inter-RAT) mobility when necessary. The control-plane functions consist of several parts. The Session Management Function (SMF) handles, among other functions, IP address allocation for the device (also known as User Equipment, UE), control of policy enforcement, and general session-management functions. The Access and Mobility Management Function (AMF) is in charge of control signaling between the core network and the device, security for user data, idle-state mobility, and authentication. The functionality operating between the core network, more specifically the AMF, and the device is sometimes referred to as the Non-Access Stratum (NAS), to separate it from the Access Stratum (AS), which handles functionality operating between the device and the radio-access network.

3.2.2 RADIO-ACCESS NETWORK

The radio-access network can have two types of nodes connected to the 5G core network: • A gNB, serving NR devices using the NR user-plane and control-plane protocols; or • An ng-eNB,

serving LTE devices using the LTE user-plane and control-plane protocols.1 A radio-access network consisting of both ng-eNBs for LTE radio access and gNBs for NR radio access is known as an NG-RAN, although the term RAN will be used in the following for simplicity. Furthermore, it will be assumed that the RAN is connected to the 5G core and hence 5G terminology, such as gNB, will be used. In other words, the description will assume a 5G core network and an NR-based RAN as shown in option 2 . However, as already mentioned, the first version of NR operates in non-standalone mode where NR is connected to the EPC using option 3. The principles are in this case similar, although the naming of the nodes and interfaces differs slightly.

The gNB (or ng-eNB) is responsible for all radio-related functions in one or several cells, for example, radio resource management, admission control, connection establishment, routing of user-plane data to the UPF and control-plane information to the AMF, and QoS flow management. It is important to note that an gNB is a logical node and not a physical implementation. One common implementation of an gNB is a three-sector site, where a base station is handling transmissions in three cells, although other implementations can be found as well, such as one baseband processing unit to which several remote radio heads are connected. Examples of the latter are a large number of indoor cells, or several cells along a highway, belonging to the same gNB. Thus, a base station is a possible implementation of, but not the same as, a gNB.

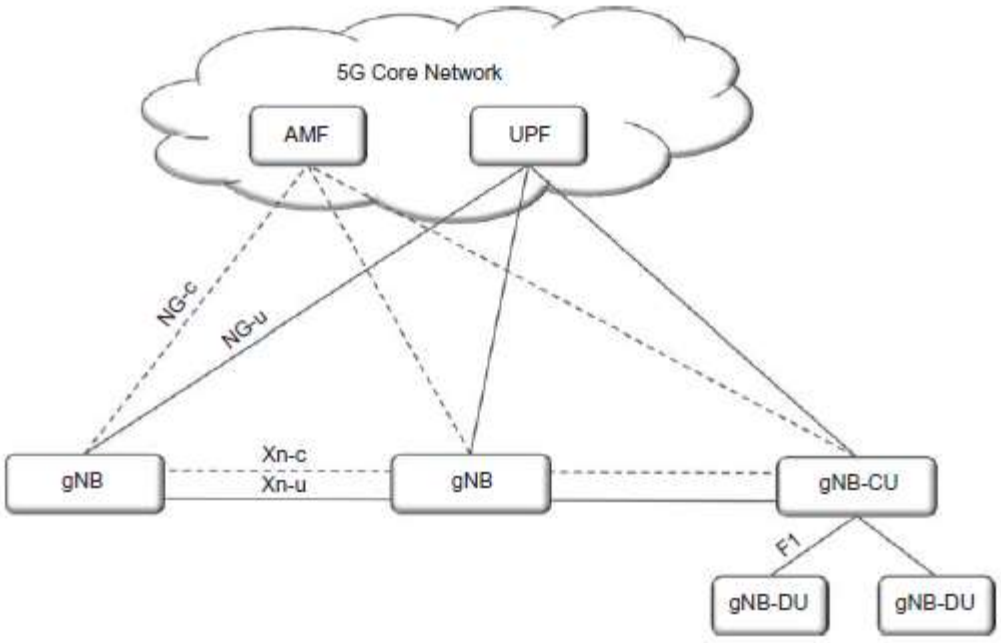


Figure 3.3 Radio-access network interfaces

The gNB is connected to the 5G core network by means of the NG interface, more specifically to the UPF by means of the NG user-plane part (NG-u), and to the AMF by means of the NG control-plane part (NG-c). One gNB can be connected to multiple UPFs/AMFs for the purpose of load sharing and redundancy. The Xn interface, connecting gNBs to each other, is mainly used

to support active-mode mobility and dual connectivity. This interface may also be used for multicell Radio Resource Management (RRM) functions. The Xn interface is also used to support lossless mobility between neighboring cells by means of packet forwarding. There is also a standardized way to split the gNB into two parts, a central unit (gNB-CU) and one or more distributed units (gNB-DU) using the F1 interface. In the case of a split gNB, the RRC, PDCP, and SDAP protocols, described in more detail below, reside in the gNB-CU and the remaining protocol entities (RLC, MAC, PHY) in the gNB-DU. The interface between the gNB (or the gNB-DU) and the device is known as the Uu interface.

3.3 QUALITY-OF-SERVICE HANDLING

Handling of different quality-of-service (QoS) requirements is possible already in LTE, and NR builds upon and enhances this framework. The key principles of LTE are kept, namely that the network is in charge of the QoS control and that the 5G core network but not the radio-access network is aware of the service. QoS handling is essential for the realization of network slicing.

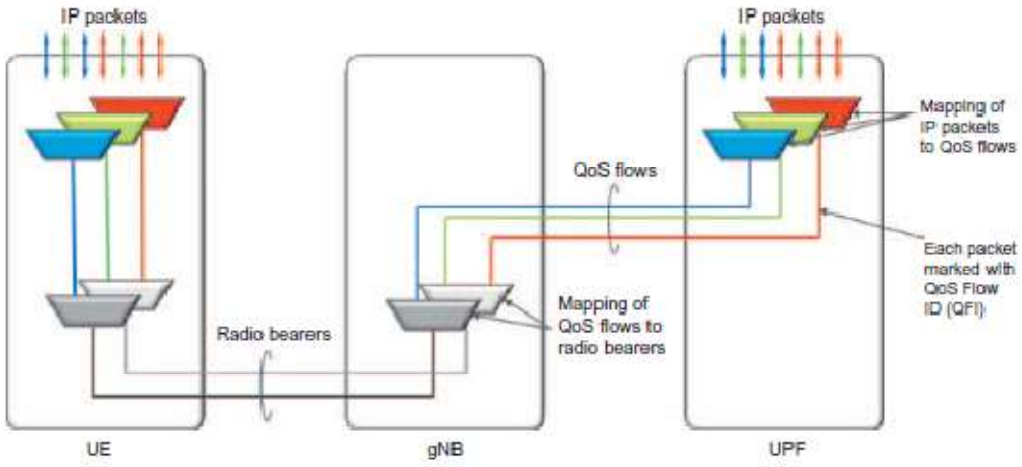


Figure 3.4 QoS flows and radio bearers during a PDU session

For each connected device, there is one or more PDU sessions, each with one or more QoS flows and data radio bearers. The IP packets are mapped to the QoS flows according to the QoS requirements, for example in terms of delay or required data rate, as part of the UDF functionality in the core network. Each packet can be marked with a QoS Flow Identifier (QFI) to assist uplink QoS handling. The second step, mapping of QoS flows to data radio bearers, is done in the radio-access network. Thus, the core network is aware of the service requirements, while the radio-access network only maps the QoS flows to radio bearers. The QoS-flow-to-radio-bearer mapping is not necessarily a one-to-one mapping; multiple QoS flows can be mapped to the same data radio bearer. There are two ways of controlling the mapping from quality-of-service flows to data radio bearers in the uplink: reflective mapping and explicit configuration. In the case of reflective mapping, which is a new feature in NR when connected to the 5G core network, the device observes the QFI in the downlink packets for the PDU session. This provides the device with knowledge about which IP flows are mapped to which QoS flow and radio bearer. The device then uses the same mapping for the uplink traffic. In the case of

explicit mapping, the quality-of-service flow to data radio bearer mapping is configured in the device using RRC signaling.

3.4 RADIO PROTOCOL ARCHITECTURE

With the overall network architecture in mind, the RAN protocol architecture for the user and control planes can be discussed.

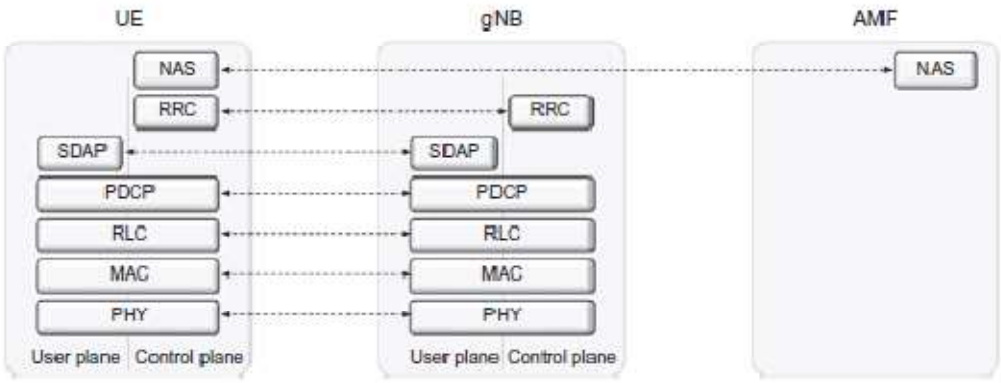


Figure 3.5 User-plane and control plane protocol stack

3.4.1 User Plane Protocols

A general overview of the NR user-plane protocol architecture for the downlink. Many of the protocol layers are similar to those in LTE, although there are some differences as well. One of the differences is the qualityof- service handling in NR when connected to a 5G core network, where the SDAP protocol layer accepts one or more QoS flows carrying IP packets according to their Quality-of-Service requirements. In the case of the NR user plane connected to the EPC, the SDAP is not used. As will become clear in the subsequent discussion, not all the entities are applicable in all situations. For example, ciphering is not used for broadcasting of the basic system information. The uplink protocol structure is similar to the downlink structure although there are some differences with respect to, for example, transport-format selection and the control of logical-channel multiplexing. The different protocol entities of the radio-access network are summarized below and described in more detail in the following sections.

• Service Data Application Protocol (SDAP) is responsible for mapping QoS bearers to radio bearers according to their quality-of-service requirements. This protocol layer is not present in LTE but introduced in NR when connecting to the 5G core network due to the new quality-of-service handling.

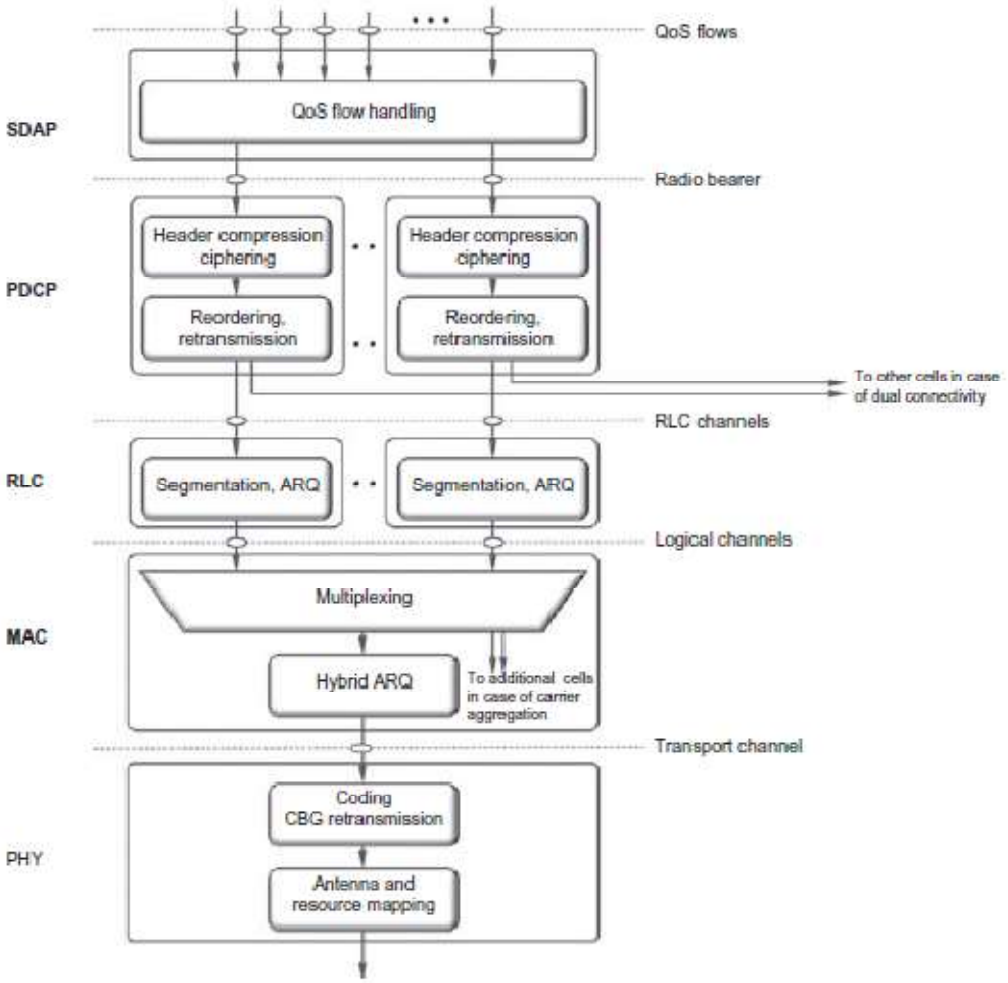


Figure 3.6 NR downlink user-plane protocol architecture as seen from the device

- Packet Data Convergence Protocol (PDCP) performs IP header compression, ciphering, and integrity protection. It also handles retransmissions, in-sequence delivery, and duplicate removal3 in the case of handover. For dual connectivity with split bearers, PDCP can provide routing and duplication. There is one PDCP entity per radio bearer configured for a device.
- Radio-Link Control (RLC) is responsible for segmentation and retransmission handling. The RLC provides services to the PDCP in the form of RLC channels. There is one RLC entity per RLC channel (and hence per radio bearer) configured for a device. Compared to LTE, the NR RLC does not support in-sequence delivery of data to higher protocol layers, a change motivated by the reduced delays as discussed below.
- Medium-Access Control (MAC) handles multiplexing of logical channels, hybrid-ARQ retransmissions, and scheduling and scheduling-related functions. The scheduling functionality is located in the gNB for both uplink and downlink. The MAC provides services to the RLC in the form of logical channels. The header structure in the MAC layer has been changed in NR to allow for more efficient support of low-latency processing than in LTE.

- Physical Layer (PHY) handles coding/decoding, modulation/demodulation, multi-antenna mapping, and other typical physical-layer functions. The physical layer offers services to the MAC layer in the form of transport channels.

3.4.2 RADIO-LINK CONTROL

The RLC protocol is responsible for segmentation of RLC SDUs from the PDCP into suitably sized RLC PDUs. It also handles retransmission of erroneously received PDUs, as well as removal of duplicate PDUs. Depending on the type of service, the RLC can be configured in one of three modes—transparent mode, unacknowledged mode, and acknowledged mode—to perform some or all of these functions. Transparent mode is, as the name suggests, transparent, and no headers are added. Unacknowledged mode supports segmentation and duplicate detection, while acknowledged mode in addition supports retransmission of erroneous packets. One major difference compared to LTE is that the RLC does not ensure insequence delivery of SDUs to upper layers. Removing in-sequence delivery from the RLC reduces the overall latency as later packets do not have to wait for retransmission of an earlier missing packet before being delivered to higher layers but can be forwarded immediately.

Another difference is the removal of concatenation from the RLC protocol to allow RLC PDUs to be assembled in advance, prior to receiving the uplink scheduling grant. This also helps reduce the overall latency. Segmentation, one of the main RLC functions, is illustrated in Fig. 6.10. Included in the figure is also the corresponding LTE functionality, which also supports concatenation. Depending on the scheduler decision, a certain amount of data, that is, certain transport-block size, is selected. As part of the overall lowlatency design of NR, the scheduling decision in case of an uplink transmission is known to the device just before transmission, in the order of a few OFDM symbols before.

In the case of concatenation in LTE, the RLC PDU cannot be assembled until the scheduling decision is known, which results in an additional delay until the uplink transmission and cannot meet the low-latency requirement of NR. By removing the concatenation from RLC, the RLC PDUs can be assembled in advance and upon receipt of the scheduling decision the device only has to forward a suitable number of RLC PDUs to the MAC layer, the number depending on the scheduled transport block size.

The RLC retransmission mechanism is also responsible for providing errorfree delivery of data to higher layers. To accomplish this, a retransmission protocol operates between the RLC entities in the receiver and transmitter. By monitoring the sequence numbers indicated in the headers of the incoming PDUs, the receiving RLC can identify missing PDUs (the RLC sequence number is independent of the PDCP sequence number). Status reports are fed back to the transmitting RLC entity, requesting retransmission of missing PDUs. Based on the received status report, the RLC entity at the transmitter can take the appropriate action and retransmit the missing PDUs if needed. Although the RLC is capable of handling transmission errors due to noise, unpredictable channel variations, etc., error-free delivery is in most cases handled by the MAC-based hybrid-ARQ protocol. The use of a retransmission mechanism in the RLC may therefore seem

superfluous at first. However, this is not the case and the use of both RLC- and MAC-based retransmission mechanisms is in fact well motivated by the differences in the feedback signaling.

3.5 MEDIUM-ACCESS CONTROL

The MAC layer handles logical-channel multiplexing, hybrid-ARQ retransmissions, and scheduling and scheduling-related functions, including handling of different numerologies. It is also responsible for multiplexing/demultiplexing data across multiple component carriers when carrier aggregation is used.

3.5.1 Logical Channels and Transport Channels

The MAC provides services to the RLC in the form of logical channels. A logical channel is defined by the type of information it carries and is generally classified as a control channel, used for transmission of control and configuration information necessary for operating an NR system, or as a traffic channel, used for the user data. The set of logical-channel types specified for NR includes:

- The Broadcast Control Channel (BCCH), used for transmission of system information from the network to all devices in a cell. Prior to accessing the system, a device needs to acquire the system information to find out how the system is configured and, in general, how to behave properly within a cell. Note that, in the case of non-standalone operation, system information is provided by the LTE system and there is no BCCH.
- The Paging Control Channel (PCCH), used for paging of devices whose location on a cell level is not known to the network. The paging message therefore needs to be transmitted in multiple cells. Note that, in the case of non-standalone operation, paging is provided by the LTE system and there is no PCCH.
- The Common Control Channel (CCCH), used for transmission of control information in conjunction with random access.
- The Dedicated Control Channel (DCCH), used for transmission of control information to/from a device. This channel is used for individual configuration of devices such as setting various parameters in devices.
- The Dedicated Traffic Channel (DTCH), used for transmission of user data to/ from a device. This is the logical channel type used for transmission of all unicast uplink and downlink user data.

The above logical channels are in general present also in an LTE system and used for similar functionality. However, LTE provides additional logical channels for features not yet supported by NR (but likely to be introduced in upcoming releases). From the physical layer, the MAC layer uses services in the form of transport channels. A transport channel is defined by how and with what characteristics the information is transmitted over the radio interface. Data on a transport channel are organized into transport blocks. In each Transmission Time Interval (TTI), at most one transport block of dynamic size is transmitted over the radio interface to/from a device (in the case of spatial multiplexing of more than four layers, there are two transport blocks per TTI).

Associated with each transport block is a Transport Format (TF), specifying how the transport block is to be transmitted over the radio interface. The transport format includes information about the transport-block size, the modulation-andcoding scheme, and the antenna mapping. By varying the transport format, the MAC layer can thus realize different data rates, a process known as transport format selection.

The following transport-channel types are defined for NR:

- The Broadcast Channel (BCH) has a fixed transport format, provided by the specifications. It is used for transmission of parts of the BCCH system information, more specifically the so-called Master Information Block (MIB)
- The Paging Channel (PCH) is used for transmission of paging information from the PCCH logical channel. The PCH supports discontinuous reception (DRX) to allow the device to save battery power by waking up to receive the PCH only at predefined time instants.
- The Downlink Shared Channel (DL-SCH) is the main transport channel used for transmission of downlink data in NR. It supports key NR features such as dynamic rate adaptation and channel-dependent scheduling in the time and frequency domains, hybrid ARQ with soft combining, and spatial multiplexing. It also supports DRX to reduce device power consumption while still providing an always-on experience. The DL-SCH is also used for transmission of the parts of the BCCH system information not mapped to the BCH. Each device has a DL-SCH per cell it is connected to. In slots where system information is received there is one additional DL-SCH from the device perspective.
- The Uplink Shared Channel (UL-SCH) is the uplink counterpart to the DLSCH— that is, the uplink transport channel used for transmission of uplink data.

In addition, the Random-Access Channel (RACH) is also defined as a transport channel, although it does not carry transport blocks. Part of the MAC functionality is multiplexing of different logical channels and mapping of the logical channels to the appropriate transport channels. The mapping between logical-channel types and transport-channel types is given. This figure clearly indicates how DL-SCH and UL-SCH are the main downlink and uplink transport channels, respectively. In the figures, the corresponding physical channels, described further below, are also included and the mapping between transport channels and physical channels is illustrated.

To support priority handling, multiple logical channels, where each logical channel has its own RLC entity, can be multiplexed into one transport channel by the MAC layer. At the receiver, the MAC layer handles the corresponding demultiplexing and forwards the RLC PDUs to their respective RLC entity. To support the demultiplexing at the receiver, a MAC header is used. The placement of the MAC headers has been improved compared to LTE, again with low-latency operation in mind. Instead of locating all the MAC header information at the beginning of a MAC PDU, which implies that assembly of the MAC PDU cannot start until the scheduling decision is available, the subheader corresponding to a certain MAC SDU is placed

immediately before the SDU. This allows the PDUs to be preprocessed before having received the scheduling decision. If necessary, padding can be appended to align the transport block size with those supported in NR.

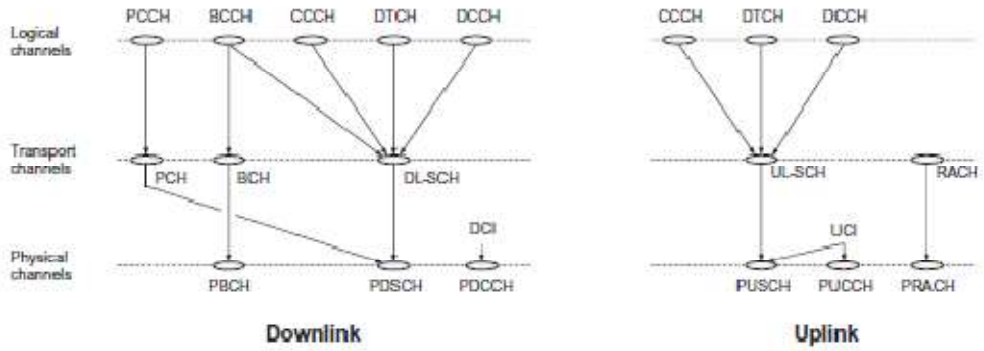


Figure 3.7 Mapping between logical, transport, and physical channels

The subheader contains the identity of the logical channel (LCID) from which the RLC PDU originated and the length of the PDU in bytes. There is also a flag indicating the size of the length indicator, as well as a reserved bit for future use. In addition to multiplexing of different logical channels, the MAC layer can also insert MAC control elements into the transport blocks to be transmitted over the transport channels. A MAC control element is used for inband control signaling and identified with reserved values in the LCID field, where the LCID value indicates the type of control information. Both fixed- and variable-length MAC control elements are supported, depending on their usage. For downlink transmissions, MAC control elements are located at the beginning of the MAC PDU, while for uplink transmissions the MAC control elements are located at the end, immediately before the padding (if present). Again, the placement is chosen in order to facilitate low-latency operation in the device. MAC control elements are, as mentioned above, used for inband control signaling. It provides a faster way to send control signaling than RLC, without having to resort to the restrictions in terms of payload sizes and reliability offered by physical-layer L1/L2 control signaling (PDCCH or PUCCH).

3.5.2 Scheduling

One of the basic principles of NR radio access is shared-channel transmission— that is, timefrequency resources are dynamically shared between users. The scheduler is part of the MAC layer (although often better viewed as a separate entity) and controls the assignment of uplink and downlink resources in terms of so-called resource blocks in the frequency domain and OFDM symbols and slots in the time domain.

The basic operation of the scheduler is dynamic scheduling, where the gNB takes a scheduling decision, typically once per slot, and sends scheduling information to the selected set of devices. Although per-slot scheduling is a common case, neither the scheduling decisions, nor the actual data transmission is restricted to start or end at the slot boundaries. This is useful to support low latency operation as well as future extensions to unlicensed spectrum operation.

Although the scheduling strategy is implementation specific and not specified by 3GPP, the overall goal of most schedulers is to take advantage of the channel variations between devices

and preferably schedule transmissions to a device on resources with advantageous channel conditions in both the time and frequency domain, often referred to as channel-dependent scheduling. Downlink channel-dependent scheduling is supported through channel-state information (CSI), reported by the device to the gNB and reflecting the instantaneous downlink channel quality in the time and frequency domains, as well as information necessary to determine the appropriate antenna processing in the case of spatial multiplexing. In the uplink, the channel-state information necessary for uplink channel-dependent scheduling can be based on a sounding reference signal transmitted from each device for which the gNB wants to estimate the uplink channel quality. To aid the uplink scheduler in its decisions, the device can transmit buffer-status and power-headroom information to the gNB using MAC control elements. This information can only be transmitted if the device has been given a valid scheduling grant. For situations when this is not the case, an indicator that the device needs uplink resources is provided as part of the uplink L1/L2 control signaling structure.

3.5.3 Hybrid ARQ with Soft Combining

Hybrid ARQ with soft combining provides robustness against transmission errors. As hybrid-ARQ retransmissions are fast, many services allow for one or multiple retransmissions, and the hybrid-ARQ mechanism therefore forms an implicit (closed loop) rate-control mechanism. The hybrid-ARQ protocol is part of the MAC layer, while the physical layer handles the actual soft combining.5 Hybrid ARQ is not applicable for all types of traffic. For example, broadcast transmissions, where the same information is intended for multiple devices, typically do not rely on hybrid ARQ. Hence, hybrid ARQ is only supported for the DL-SCH and the UL-SCH, although its usage is up to the gNB implementation.

The hybrid-ARQ protocol uses multiple parallel stop-and-wait processes in a similar way to LTE. Upon receipt of a transport block, the receiver tries to decode the transport block and informs the transmitter about the outcome of the decoding operation through a single acknowledgment bit indicating whether the decoding was successful or if a retransmission of the transport block is required. Clearly, the receiver must know to which hybrid-ARQ process a received acknowledgment is associated. This is solved by using the timing of the acknowledgment for association with a certain hybrid-ARQ process or by using the position of the acknowledgment in the hybrid-ARQ codebook in case of multiple acknowledgments transmitted at the same time.

Up to 16 hybrid-ARQ processes are supported. Having a larger maximum number of hybrid-ARQ processes than in LTE7 is motivated by the possibility for remote radio heads, which incurs a certain front-haul delay, together with the shorter slot durations at high frequencies. It is important though, that the larger number of maximum hybrid-ARQ processes does not imply a longer roundtrip time as not all processes need to be used, it is only an upper limit of the number of processes possible. The use of multiple parallel hybrid-ARQ processes, for a device can result in data being delivered from the hybrid-ARQ mechanism out of sequence. For example, transport block 3 in the figure was successfully decoded before transport block 2, which required retransmissions.

3.6 PHYSICAL LAYER

The physical layer is responsible for coding, physical-layer hybrid-ARQ processing, modulation, multi-antenna processing, and mapping of the signal to the appropriate physical timefrequency resources. It also handles mapping of transport channels to physical channels, as shown in Fig. 6.11. As mentioned in the introduction, the physical layer provides services to the MAC layer in the form of transport channels. Data transmissions in downlink and uplink use the DL-SCH and UL-SCH transport-channel types, respectively. There is at most one transport block (two transport blocks in the case of spatial multiplexing of more than four layers in the downlink) to a single device per TTI on a DL-SCH or UL-SCH. In the case of carrier aggregation, there is one DL-SCH (or UL-SCH) per component carrier seen by the device. A physical channel corresponds to the set of timefrequency resources used for transmission of a particular transport channel and each transport channel is mapped to a corresponding physical channel. In addition to the physical channels with a corresponding transport channel, there are also physical channels without a corresponding transport channel. These channels, known as L1/L2 control channels, are used for downlink control information (DCI), providing the device with the necessary information for proper reception and decoding of the downlink data transmission, and uplink control information (UCI) used for providing the scheduler and the hybrid-ARQ protocol with information about the situation at the device.

The following physical-channel types are defined for NR:

• The Physical Downlink Shared Channel (PDSCH) is the main physical channel used for unicast data transmission, but also for transmission of, for example, paging information, random-access response messages, and delivery of parts of the system information.

• The Physical Broadcast Channel (PBCH) carries part of the system information, required by the device to access the network.

• The Physical Downlink Control Channel (PDCCH) is used for downlink control information, mainly scheduling decisions, required for reception of PDSCH, and for scheduling grants enabling transmission on the PUSCH.

• The Physical Uplink Shared Channel (PUSCH) is the uplink counterpart to the PDSCH. There is at most one PUSCH per uplink component carrier per device.

• The Physical Uplink Control Channel (PUCCH) is used by the device to send hybrid-ARQ acknowledgments, indicating to the gNB whether the downlink transport block(s) was successfully received or not, to send channel-state reports aiding downlink channel-dependent scheduling, and for requesting resources to transmit uplink data upon.

• The Physical Random-Access Channel (PRACH) is used for random access.

Note that some of the physical channels, more specifically the channels used for downlink and uplink control information (PDCCH and PUCCH) do not have a corresponding transport channel mapped to them.

3.7 CONTROL-PLANE PROTOCOLS

The control-plane protocols are, among other things, responsible for connection setup, mobility, and security. The NAS control-plane functionality operates between the AMF in the core network and the device. It includes authentication, security, and different idle mode procedures such as paging (described below). It is also responsible for assigning an IP address to a device. The Radio Resource Control (RRC) control-plane functionality operates between the

RRC located in the gNB. RRC is responsible for handling the RANrelated control-plane procedures, including:

Broadcast of system information necessary for the device to be able to communicate with a cell. Acquisition of system information.

• Transmission of paging messages originating from the MME to notify the device about incoming connection requests. Paging is used in the RRC_IDLE state (described further below) when the device is not connected to a cell. Indication of system-information updates is another use of the paging mechanism, as is public warning systems.

• Connection management, including setting up bearers and mobility. This includes establishing an RRC context—that is, configuring the parameters necessary for communication between the device and the radio-access network.

• Mobility functions such as cell (re)selection.

• Measurement configuration and reporting.

• Handling of device capabilities; when connection is established the device will announce its capabilities as not all devices are capable of supporting all the functionality described in the specifications. RRC messages are transmitted to the device using signaling radio bearers (SRBs), using the same set of protocol layers (PDCP, RLC, MAC, and PHY). The SRB is mapped to the common control channel (CCCH) during establishment of connection and, once a connection is established, to the dedicated control channel (DCCH). Control-plane and user-plane data can be multiplexed in the MAC layer and transmitted to the device in the same TTI. The aforementioned MAC control elements can also be used for control of radio resources in some specific cases where low latency is more important than ciphering, integrity protection, and reliable transfer.

3.7.1 RRC STATE MACHINE

In most wireless communication systems, the device can be in different states depending on the traffic activity. This is true also for NR and an NR device can be in one of three RRC states, RRC_IDLE, RRC_ACTIVE, and RRC_INACTIVE. The first two RRC states, RRC_IDLE and RRC_CONNECTED, are similar to the counterparts in LTE, while RRC_INACTIVE is a new state introduced in NR and not present in the original LTE design. There are also core network states not discussed further herein, CN_IDLE and CN_CONNECTED, depending on whether the device has established a connection with the core network or not.

Figure 3.8 RRC states

In RRC_IDLE, there is no RRC context—that is, the parameters necessary for communication between the device and the network—in the radio-access network and the device does not belong to a specific cell. From a core network perspective, the device is in the CN_IDLE state. No data transfer may take place as the device sleeps most of the time to reduce battery consumption. In

the downlink, devices in idle state periodically wake up to receive paging messages, if any, from the network. Mobility is handled by the device through cell reselection. Uplink synchronization is not maintained and hence the only uplink transmission activity that may take place is random access, to move to a connected state. As part of moving to a connected state, the RRC context is established in both the device and the network. In RRC_CONNECTED, the RRC context is established and all parameters necessary for communication between the device and the radio-access network are known to both entities. From a core network perspective, the device is in the CN_CONNECTED state. The cell to which the device belongs is known and an identity of the device, the Cell Radio-Network Temporary Identifier (C-RNTI), used for signaling purposes between the device and the network, has been configured. The connected state is intended for data transfer to/from the device, but discontinuous reception (DRX) can be configured to reduce device power consumption. Since there is an RRC context established in the gNB in the connected state, leaving DRX and starting to receive/transmit data is relatively fast as no connection setup with its associated signaling is needed. Mobility is managed by the radio-access network, that is, the device provides neighboring-cell measurements to the network which commands the device to perform a handover when relevant. Uplink time alignment may or may not exist but need to be established using random access and maintained as described in Section 16.2 for data transmission to take place. In LTE, only idle and connected states are supported. A common case in practice is to use the idle state as the primary sleep state to reduce the device power consumption. However, as frequent transmission of small packets is common for many smartphone applications, the result is a significant amount of idle-to-active transitions in the core network. These transitions come at a cost in terms of signaling load and associated delays. Therefore, to reduce the signaling load and in general reduce the latency, a third state is defined in NR, the RRC_INACTIVE state.

In RRC_INACTIVE, the RRC context is kept in both the device and the gNB. The core network connection is also kept, that is, the device is in CN_CONNECTED from a core network perspective. Hence, transition to connected state for data transfer is fast. No core network signaling is needed. The RRC context is already in place in the network and idle-to-active transitions can be handled in the radio-access network. At the same time, the device is allowed to sleep in a similar way as in the idle state and mobility is handled through cell reselection, that is, without involvement of the network. Thus, RRC_INACTIVE can be seen as a mix of the idle and connected states.As seen from the discussion above, one important difference between the different states is the mobility mechanisms involved. Efficient mobility handling is a key part of any mobile communication system. For the idle and inactive states, mobility is handled by the device through cell reselection, while for the connected mode, mobility is handled by the radio-access network based on measurements. The different mobility mechanisms are described below, starting with idle- and inactive-mode mobility.

Idle-State And Inactive-State Mobility

The purpose of the mobility mechanism in idle and inactive states is to ensure that a device is reachable by the network. The network does this by notifying the device by means of a paging message. The area over which such a paging message is transmitted is a key aspect of the paging mechanism and in idle and inactive modes, the device is in control on when to update this

information. This is sometimes referred to as cell reselection. In essence, the device searches for and measures on candidate cells similar to the initial cell search. Once the device discovers a cell with a received power sufficiently higher than its current one, it considered this as the best cell and, if necessary, contacts the network through random access.

Connected-State Mobility

In a connected state the device has a connection established to the network. The aim of connected-state mobility is to ensure that this connectivity is retained without any interruption or noticeable degradation as the device moves within the network. To ensure this, the device continuously searches for new cells both on the current carrier frequency (intra-frequency measurements) and on different carrier frequencies (inter-frequency measurements) that the device has been informed about. Such measurements can be done on an SS block in essentially the same way as for initial access and cell search in idle and inactive mode. However, measurements can also be done on configured CSI-RS. In a connected state, the device does not make any decisions of its own when it comes to handover to a different cell. Rather, based on different triggering conditions, for example, the relative power of a measured SS block compared to the current cell, the device reports the result of the measurements to the network. Based on this reporting the network makes a decision as to whether or not the device is to handover to a new cell. It should be pointed out that this reporting is done using RRC signaling, that is, it is not covered by the Layer-1 measurement and reporting framework used, for example, for beam management. Except for very small cells that are tightly synchronized to each other, the current uplink transmission timing of a device will typically not match the new cell to which a device is assumed to handover. To establish synchronization to a new cell a device thus has to carry out a procedure similar to the random-access procedure. However, this may then be a contention-free random access using resources specifically assigned to the device with no risk for collision but only aiming at establishing synchronization to the new cell. Thus, only the two first steps of the random-access procedure are needed, that is, the preamble transmission and corresponding random-access response providing the device with updated transmission timing.

Reference Books:

1. Saad Z. Asif, "5G Mobile Communications Concepts and Technologies, CRC Press, 1st Edition, 2019.
2. Erik Dahlman, Stefan Parkvall, Johan Skold "5G NR: The Next Generation Wireless Access Technology", Academic Press, 1st Edition, 2018.
3. Jonathan Rodriguez, "Fundamentals 5G Mobile Networks", John Wiley & Sons, 1st Edition, 2015.
4. Long Zhao, Hui Zhao, Kan Zheng, Wei Xiang, "Massive MIMO in 5G Networks: Selected Applications", Springer, 1st Edition, 2018.
5. Robert W. Heath Jr., Angel Lozano, "Foundations of MIMO Communication", Cambridge University Press, 1st Edition, 2019.

6. R. Vannithamby and S. Talwar, "Towards 5G: Applications, Requirements and Candidate Technologies", John Willey & Sons, 1st Edition, 2017.

Questions to Practice:

PART -A

1	Sketch the 5G NR Requirements
2	Interpret Ultra Lean Design on 5G Communication
3	Contrast Service Data Adaption Protocol in detail
4	Relate how Scheduling is differ from previous Technologies
5	Examine Control Channels in 5G NR

PART-B

1	Demonstrate in detail about the design Principles of 5G NR
2	Implement an Overall System Architecture to overcome the Poor phase synchronization
3	Contrast User Plane Protocols in 5g NR which can overcome the issues.
4	Examine in detail about Medium Access Control in 5g NR

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF ELECTRICAL AND ELECTRONICS

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

UNIT – IV –CHANNEL MODELS FOR 5G NR– SECA3020

UNIT-IV (CHANNEL MODELS FOR 5G NR)

Channel Hierarchy in 5G NR – Logical Channels and Transport Channels in 5G NR - Physical Layer Data Channels in 5G NR - Downlink Physical Channel and Uplink Physical Channels - Propagation Channel models for 5G

4.1 Introduction :

The physical layer provides services to the MAC layer in the form of transport channels as described in Section 6.4.5. In the downlink, there are three different types of transport channels defined for NR: the Downlink Shared Channel (DLSCH), the Paging Channel (PCH), and the Broadcast Channel (BCH), although the latter two are not used in the non-standalone operation. In the uplink, there is only one uplink transport-channel type carrying transport blocks in NR,1 the Uplink Shared Channel (UL-SCH). The overall transport channel processing for NR follows a similar structure as for LTE.

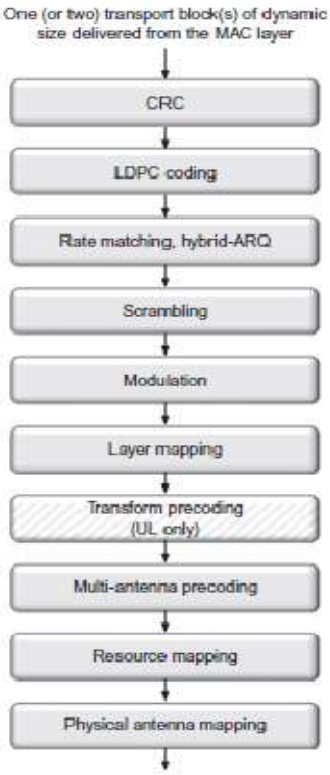


Figure 4.1: General transport Channel Processing

The processing is mostly similar in uplink and downlink and the structure is applicable for the DL-SCH, BCH, and PCH in the downlink, and the UL-SCH in the uplink. The part of the BCH that is mapped to the PBCH follows a different structure, described in Section 16.1, as does the RACH. Within each transmission time interval (TTI), up to two transport blocks of dynamic size are delivered to the physical layer and transmitted over the radio interface for each component carrier. Two transport blocks are only used in the case of spatial multiplexing with more than four layers, which is only supported in the downlink direction and mainly useful in scenarios with very high signal-tonoise ratios. Hence, at most a single transport block per component carrier and TTI is a typical case in practice. A CRC for error-detecting purposes is added to each transport block, followed by error-correcting coding using LDPC codes. Rate matching, including physicallayer hybrid-ARQ functionality, adapts the number of coded bits to the scheduled resources. The code bits are scrambled and fed to a modulator, and finally the modulation symbols are mapped to the physical resources, including the spatial domain. For the uplink there is also a possibility of a DFT-precoding. The differences between uplink and downlink is, apart from DFT-precoding being possible in the uplink only, mainly around antenna mapping and associated reference signals.

4.2 CHANNEL CODING:

An overview of the channel coding steps is provided and described in more detail in the following sections. First, a CRC is attached to the transport block to facilitate error detection, followed by code block segmentation. Each code block is LDPC-encoded and rate matched separately, including physicallayer hybrid-ARQ processing, and the resulting bits are concatenated to form the sequence of bits representing the coded transport block.

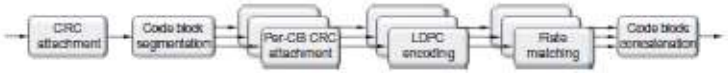


Figure 4.2 Channel Coding

4.2.1 CRC ATTACHMENT PER TRANSPORT BLOCK: In the first step of the physical-layer processing, a CRC is calculated for and appended to each transport block. The CRC allows for receiver-side detection of errors in the decoded transport block and can, for example, be used by the hybrid-ARQ protocol as a trigger for requesting retransmissions. The size of the CRC depends on the transport-block size. For transport blocks larger than 3824 bits, a 24-bit CRC is used, otherwise a 16-bit CRC is used to reduce overhead.

4.2.2 CODE-BLOCK SEGMENTATION: The LDPC coder in NR is defined up to a certain code-block size (8424 bits for base graph 1 and 3840 bits for base graph 2). To handle transport block sizes larger than this, code-block segmentation is used where the transport block, including the CRC, is split into multiple equal-sized2 code blocks as illustrated.

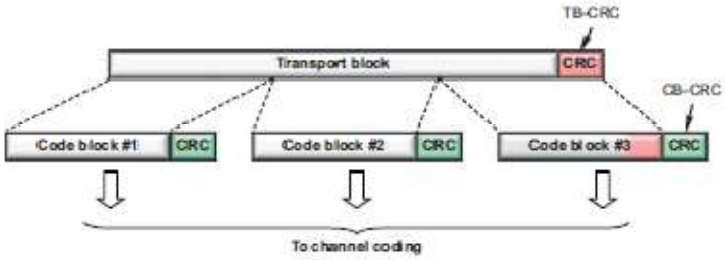


Figure 4.3 Code Block Segmentation

Code-block segmentation also implies that an additional CRC (also of length 24 bits but different compared to the transport-block CRC described above) is calculated for and appended to each code block. In the case of a single code-block transmission no additional code-block CRC is applied. One could argue that, in the case of code-block segmentation, the transport block CRC is redundant and implies unnecessary overhead as the set of codeblock CRCs should indirectly provide information about the correctness of the complete transport block. However, to handle code-block group (CBG) retransmissions as discussed in Chapter 13, a mechanism to detect errors per code block is necessary. CBG retransmission means that only the erroneous code-block groups are retransmitted instead of the complete transport block to improve the spectral efficiency. The per-CB CRC can also be used for the device to limit decoding in case of a retransmission only to those CBs whose CRCs did not check even if per-CBG retransmission is not configured. This helps reducing the device processing load. The transport-block CRC also adds an extra level of protection in terms of error detection. Note that code-block segmentation is only applied to large transport blocks for which the relative extra overhead due to the additional transport-block CRC is small.

4.3 CHANNEL CODING:

Channel coding is based on LDPC codes, a code design which was originally proposed in the 1960s but forgotten for many years. They were “rediscovered” in the 1990s and found to be an attractive choice from an implementation perspective. From an error-correcting capability point of view, turbo codes, as used in LTE, can achieve similar performance, but LDPC codes can offer lower complexity, especially at higher code rates, and were therefore chosen for NR. The basis for LDPC codes is a sparse (low-density) parity check matrix H where for each valid code word c the relation HcT50 holds. Designing a good LDPC code to a large extent boils down to finding a good parity check matrix H which is sparse (the sparseness implies relatively simple decoding). It is common to represent the parity-check matrix by a graph connecting n variable nodes at the top with (nk) constraint nodes at the bottom of the graph, a notation that allows a wide range of properties of an (n, k) LDPC code to be analyzed. This explains why the term base graph is used in the NR specifications. A detailed description of the theory behind LDPC codes is beyond the scope of this book, but there is a rich literature in the field (for example, see [68]). Quasi-cyclic LDPC codes with a dual-diagonal structure of the kernel part of the parity check matrix are used in NR, which gives a decoding complexity which is linear in the number of coded bits and enables a simple encoding operation. Two base graphs are defined, BG1 and BG2, representing the two base matrices. The reason for two base graphs

instead of one is to handle the wide range of payload sizes and code rates in an efficient way. Supporting a very large payload size at a medium to high code rate, which is the case for very high data rates, using a code designed to support a very low code rate is not efficient. At the same time, the lowest code rates are necessary to provide good performance in challenging situations. In NR, BG1 is designed for code rates from 1/3 to 22/24 (approximately 0.330.92) and BG 2 from 1/5 to 5/6 (approximately 0.20.83). Through puncturing, the highest code rate can be increased somewhat, up to 0.95, beyond which the device is not required to decode. The choice between BG1 and BG2 is based on the transport block size and code rate targeted for the first transmission.

In short, for a given lifting size Z, each "1" in the base matrix is replaced by the Z3Z identify matrix circularly shifted by the corresponding shift coefficient and each "0" in the base matrix is replaced by the Z3Z all-zero matrix. Hence, a relatively large number of parity-check matrices can be generated to support multiple payload sizes while maintaining the general structure of the LDPC code. To support payload sizes that are not a native payload size of one of the 51 defined parity check matrices, known filler bits can be appended to the code block before encoding. Since the NR LDPC codes are systematic codes, the filler bits can be removed before transmission.

4.4 SCRAMBLING

Scrambling is applied to the block of coded bits delivered by the hybrid-ARQ functionality by multiplying the sequence of coded bits with a bit-level scrambling sequence. Without scrambling, the channel decoder at the receiver could, at least in principle, be equally matched to an interfering signal as to the target signal, thus being unable to properly suppress the interference. By applying different scrambling sequences for neighboring cells in the downlink or for different devices in the uplink, the interfering signal(s) after descrambling is (are) randomized, ensuring full utilization of the processing gain provided by the channel code. The scrambling sequence in both downlink (PDSCH) and uplink (PUSCH) depends on the identity of the device, that is, the C-RNTI, and a data scrambling identity configured in each device. If no data scrambling identity is configured, the physical layer cell identity is used as a default value to ensure that neighboring devices, both in the same cell and between cells, use different scrambling sequences. Furthermore, in the case of two transport blocks being transmitted in the downlink to support more than four layers, different scrambling sequences are used for the two transport blocks.

4.5 MODULATION

The modulation step transforms the block of scrambled bits to a corresponding block of complex modulation symbols. The modulation schemes supported include QPSK, 16QAM, 64QAM, and 256QAM in both uplink and downlink. In addition, for the uplink π/2-BPSK is supported in the case the DFT-precoding is used, motivated by a reduced cubic metric [60] and hence improved power amplifier efficiency, in particular for coverage limited scenarios. Note that π/2- BPSK is neither supported nor useful in the absence of DFT-precoding as the cubic metric in this case is dominated by the OFDM waveform.

4.6 LAYER MAPPING

The purpose of the layer-mapping step is to distribute the modulation symbols across the different transmission layers. This is done in a similar way as for LTE; every nth symbol is mapped to the nth layer. One coded transport block can be mapped on up to four layers. In the case of five to eight layers, supported in the downlink only, a second transport block is mapped to layers five to eight following the same principle as for the first transport block.

Multi-layer transmission is only supported in combination with OFDM, the baseline waveform in NR. With DFT-precoding in the uplink, only a single transmission layer is supported. This is motivated both by the receiver complexity, which in the case of multi-layer transmission would be significantly higher with a DFT-precoder than without, and the use case originally motivating the additional support of DFT-precoding, namely handling of coverage-limited scenarios. In such a scenario, the received signal-to-noise ratio is too low for efficient usage of spatial multiplexing and there is no need to support spatial multiplexing to a single device.

4.7 Physical-Layer Control Signaling

To support the transmission of downlink and uplink transport channels, there is a need for certain associated control signaling. This control signaling is often referred to as L1/L2 control signaling, indicating that the corresponding information partly originates from the physical layer (layer 1) and partly from MAC (layer 2).

4.7.1 Downlink:

Downlink L1/L2 control signaling consists of downlink scheduling assignments, including information required for the device to be able to properly receive, demodulate, and decode the DL-SCH on a component carrier, and uplink scheduling grants informing the device about the resources and transport format to use for uplink (ULSCH) transmission. In addition, the downlink control signaling can also be used for special purposes such as conveying information about the symbols used for uplink and downlink in a set of slots, preemption indication, and power control. In NR, there is only a single control channel, the physical downlink control channel (PDCCH). On a high level, the principles of the PDCCH processing in NR are similar to LTE, namely that the device tries to blindly decode candidate PDCCHs transmitted from the network using one or more search spaces. However, there are some differences compared to LTE based on the different design targets for NR as well as experience from LTE deployments: • The PDCCH in NR does not necessarily span the full carrier bandwidth, unlike the LTE PDCCH. This is a natural consequence of the fact that not all NR devices may be able to receive the full carrier bandwidth, and led to the design of a more generic control channel structure in NR. • The PDCCH in NR is designed to support device-specific beamforming, in line with the general beam-centric design of NR and a necessity when operating at very high carrier frequencies with a corresponding challenging link budget.

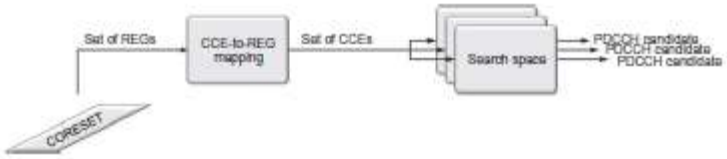


Figure 4.4 Overview of PDCCH processing in NR

These two aspects were to some extent addressed in the LTE EPDCCH design in release 11, although in practice EPDCCH has not been used extensively except as a basis for the control signaling for eMTC. Two other control channels present in LTE, the PHICH and the PCFICH, are not needed in NR. The former is used in LTE to handle uplink retransmissions and is tightly coupled to the use of a synchronous hybrid-ARQ protocol, but since the NR hybrid-ARQ protocol is asynchronous in both uplink and downlink the PHICH is not needed in NR. The latter channel, the PCFICH, is not necessary in NR as the size of the control resource sets (CORESETs) does not vary dynamically and reuse of control resources for data is handled in a different way than in LTE, as discussed further below.

In the following sections, the NR downlink control channel, the PDCCH, will be described, including the notion of a CORESETs, the timefrequency resources upon which the PDCCH is transmitted. First, the PDCCH processing including coding and modulation will be discussed, followed by a discussion on the CORESETs structure. There can be multiple CORESETs on a carrier and part of the control resource set is the mapping from resource elements to control channel elements (CCEs). One or more CCEs from one control resource set are aggregated to form the resources used by one PDCCH. Blind detection, the process where the device attempts to detect if there are any PDCCHs transmitted to the device, is based on search spaces. There can be multiple search spaces using the resources in a single CORESET, as illustrated in Fig. 10.1. Finally, the contents of the downlink control information (DCI) will be described.

Physical Downlink Control Channel:

The PDCCH processing steps are illustrated. At a high level, the PDCCH processing in NR is more similar to the LTE EPDCCH than the LTE PDCCH in the sense that each PDCCH is processed independently. The payload transmitted on a PDCCH is known as Downlink Control Information (DCI) to which a 24-bit CRC is attached to detect transmission errors and to aid the decoder in the receiver. Compared to LTE, the CRC size has been increased to reduce the risk of incorrectly received control information and to assist early termination of the decoding operation in the receiver.

Similarly to LTE, the device identity modifies the CRC transmitted through a scrambling operation. Upon receipt of the DCI, the device will compute a scrambled CRC on the payload part using the same procedure and compare it against the received CRC. If the CRC checks, the message is declared to be correctly received and intended for the device. Thus, the identity of the device that is supposed to receive the DCI message is implicitly encoded in the CRC and not explicitly transmitted. This reduces the number of bits necessary to transmit on the PDCCH as, from a device point of view, there is no difference between a corrupt message whose CRC will

not check, and a message intended for another device. Note that the RNTI does not necessarily have to be the identity of the device, the C-RNTI, but can also be different types of group or common RNTIs, for example, to indicate paging or a random-access response.

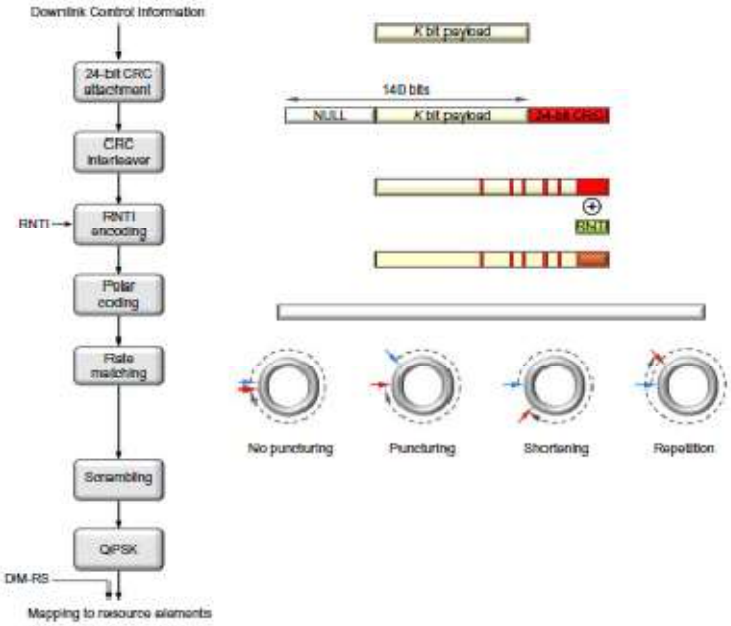


Figure 4.5 PDCCH Processing

Channel coding of the PDCCH is based on Polar codes, a relatively new form of channel coding. The basic idea behind Polar codes is to transform several instances of the radio channel into a set of channels that are either noiseless or completely noisy and then transmit the information bits on the noiseless channels. Decoding can be done in several ways, but a typical approach is to use successive cancellation and list decoding. List decoding uses the CRC as part of the decoding process, which means that the error-detecting capabilities are reduced. For example, list decoding of size eight results in a loss of three bits from an errordetecting perspective, resulting in the 24-bits CRC providing error-detecting capabilities corresponding to a 21-bit CRC. This is part of the reason for the larger CRC size compared to LTE. Unlike the tailbiting convolutional codes used in LTE, which can handle any number of information bits, Polar codes need to be designed with a maximum number of bits in mind. In NR, the Polar code has been designed to support 512 coded bits (prior to rate matching) in the downlink. Up to 140 information bits can be handled, which provides a sufficient margin for future extensions as the DCI payload size in release 15 is significantly less. To assist early termination in the decoding process, the CRC is not attached at the end of the information bits, but inserted in a distributed manner, after which the Polar code is applied.

Early termination can also be achieved by exploiting the path metric in the decoder. Rate matching is used to match the number of coded bits to the resources available for PDCCH transmission. This is a somewhat intricate process and is based on one of shortening, puncturing, or repetition of the coded bits after subblock interleaving of 32 blocks. The set of

rules selecting between shortening, puncturing, and repetition, as well as when to use which of the schemes, is designed to maximize performance. Finally, the coded and rate-matched bits are scrambled, modulated using QPSK, and mapped to the resource elements used for the PDCCH, the details of which will be discussed below. Each PDCCH has its own reference signal, which means that the PDCCH can make full use of the antenna setup, for example, be beamformed in a particular direction.

Control Resource Set:

Central to downlink control signaling in NR is the concept of CORESETs. A control resource set is a time frequency resource in which the device tries to decode candidate control channels using one or more search spaces. The size and location of a CORESET in the timefrequency domain is semistatically configured by the network and can thus be set to be smaller than the carrier bandwidth. This is especially important in NR as a carrier can be very wide, up to 400 MHz, and it is not reasonable to assume all devices can receive such a wide bandwidth. In LTE, the concept of a CORESET is not explicitly present. Instead, downlink control signaling in LTE uses the full carrier bandwidth in the first 1-3 OFDM symbols (four for the most narrowband case). This is known as the control region in LTE and in principle this control region would correspond to the "LTE CORESET" if that term would have been used. Having the control channels spanning the full carrier bandwidth was well motivated by the desire for frequency diversity and the fact that all LTE devices support the full 20 MHz carrier bandwidth (at least at the time of specifying release 8). However, in later LTE releases this lead to complications when introducing support for devices not supporting the full carrier bandwidth, for example, the eMTC devices introduced in release 12. Another drawback of the LTE approach is the inability to handle frequency domain interference coordination between cells for the downlink control channels. To some extent, these drawbacks with the LTE control channel design were addressed with the introduction of the EPDCCH in release 11, but the EPDCCH feature has so far not been widely deployed in practice as an LTE network still needs to provide PDCCH support for initial access and to handle non-EPDCCH capable LTE devices. Therefore, a more flexible structure is used in NR from the start.

A CORESET can occur at any position within a slot and anywhere in the frequency range of the carrier. However, a device is not expected to handle CORESETs outside its active bandwidth part. The reason for configuring CORESETs on the cell level and not per bandwidth part is to facilitate reuse of CORSETs between bandwidth parts, for example, when operating with bandwidth adaptation.

The first CORSET, CORESET 0, is provided by the master information block (MIB) as part of the configuration of the initial bandwidth part to be able to receive the remaining system information and additional configuration information from the network. After connection setup, a device can be configured with multiple, potentially overlapping, CORESETs in addition to using RRC signaling.

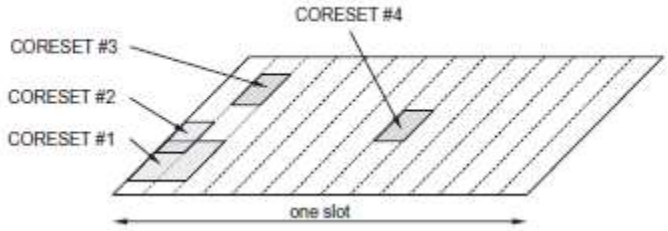


Figure 4.6 CORESET Configurations

In the time domain, a CORESET can be up to three OFDM symbols in duration and located anywhere within a slot, although a common scenario, suitable for traffic scenarios when a scheduling decision is taken once per slot, is to locate the CORESET at the beginning of the slot. This is similar to the LTE situation with control channels at the beginning of each LTE subframe. However, configuring a CORESET at other time instances can be useful, for example to achieve very low latency for transmissions occupying only a few OFDM symbols without waiting for the start of the next slot. It is important to understand that a CORESET is defined from a device perspective and only indicates where a device may receive PDCCH transmissions. It does not say anything on whether the gNB actually transmits a PDCCH or not. Depending on where the front-loaded DM-RS for PDSCH are located, in the third or fourth OFDM symbol of a slot, the maximum duration for a CORESET is two or three OFDM symbols. This is motivated by the typical case of locating the CORESET before the start of downlink reference signals and the associated data. In the frequency domain, a CORESET is defined in multiples of six resource blocks up to the carrier bandwidth. Unlike LTE, where the control region can vary dynamically in length as indicated by a special control channel (the PCFICH), a CORESET in NR is of fixed size. This is beneficial from an implementation perspective, both for the device and the network. From a device perspective, a pipelined implementation is simpler if the device can directly start to process the PDCCH without having to first decode another channel like the PCFICH in LTE. Having a streamlined and implementation-friendly structure of the PDCCH is important in order to realize the very low latency possible in NR. However, from a spectral efficiency point of view, it is beneficial if resources can be shared flexibly between control and data in a dynamic manner. Therefore, NR provides the possibility to start the PDSCH data before the end of a CORESET. It is also possible to, for a given device, reuse unused CORESET resources.

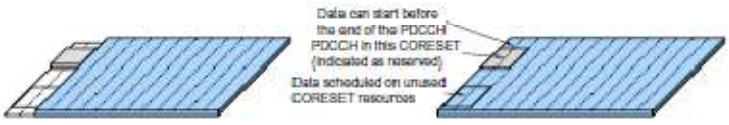


Figure 4.7 No reuse (left) and reuse (right) of CORESET resources for data transmission

4.7.2 Uplink:

Similar to LTE, there is also a need for uplink L1/L2 control signaling to support data transmission on downlink and uplink transport channels. Uplink L1/L2 control signaling consists of: • Hybrid-ARQ acknowledgments for received DL-SCH transport blocks; • Channel-state information (CSI) related to the downlink channel conditions, used to assist downlink scheduling, including multi-antenna and beamforming schemes; and • Scheduling requests, indicating that a device needs uplink resources for ULSCH transmission.

There is no UL-SCH transport-format information included in the uplink transmission. As mentioned in Section 6.4.4, the gNB is in complete control of the uplink UL-SCH transmissions and the device always follows the scheduling grants received from the network, including the UL-SCH transport format specified in those grants. Thus, the network knows the transport format used for the UL-SCH transmission in advance and there is no need for any explicit transportformat signaling on the uplink. The physical uplink control channel (PUCCH) is the basis for transmission of uplink control. In principle, the UCI could be transmitted on the PUCCH regardless of whether the device is transmitting data on the PUSCH. However, especially if the uplink resources for the PUSCH and the PUCCH are on the same carrier (or, to be more precise, use the same power amplifier) but widely separated in the frequency domain, the device may need a relatively large power back-off to fulfill the spectral emission requirements with a corresponding impact on the uplink coverage. Hence, similarly to LTE, NR supports UCI on PUSCH as the basic way of handling simultaneous transmission of data and control. Thus, if the device is transmitting on the PUSCH the UCI is multiplexed with data on the granted resources instead of being transmitted on the PUCCH. Simultaneous PUSCH and PUCCH is not part of release 15 but may be introduced in a later release.

Beamforming can be applied to the PUCCH. This is realized by configuring one or more spatial relations between the PUCCH and downlink signals such as CSI-RS or SS block. In essence, such a spatial relation means that the device can transmit the uplink PUCCH using the same beam as it used for receiving the corresponding downlink signal. For example, if the spatial relation between PUCCH and SS block is configured, the device will transmit PUCCH using the same beam as it used for receiving the SS block. Multiple spatial relations can be configured and MAC control elements used to indicate which one to use. In the case of carrier aggregation, the uplink control information is transmitted on the primary cell as a baseline. This is motivated by the need to support asymmetric carrier aggregation with the number of downlink carriers supported by a device that is unrelated to the number of uplink carriers. For a large number of downlink component carriers, a single uplink carrier may carry a large number of acknowledgments. To avoid overloading a single carrier, it is possible to configure two PUCCH groups where feedback relating to the first group is transmitted in the uplink of the PCell and feedback relating to the other group of carriers is transmitted on the primary second cell (PSCell). In the following section, the basic PUCCH structure and the principles for PUCCH control signaling are described, followed by control signaling on PUSCH.

4.7.2.1 Basic PUCCH Structure

Uplink control information can be transmitted on PUCCH using several different formats.

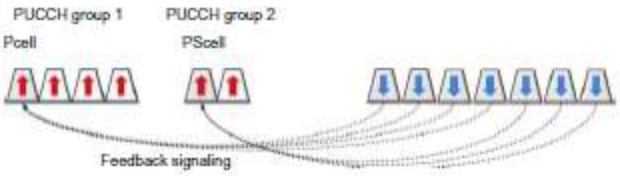


Figure 4.8 PUCCH Groups

Two of the formats, 0 and 2, are sometimes referred to as short PUCCH formats, as they occupy at most two OFDM symbols. In many cases the last one or two OFDM symbols in a slot are used for PUCCH transmission, for example, to transmit a hybrid-ARQ acknowledgment of the downlink data transmission. The short PUCCH formats include:

• PUCCH format 0, capable of transmitting at most two bits and spanning one or two OFDM symbols. This format can, for example, be used to transmit a hybrid-ARQ acknowledgment of a downlink data transmission, or to issue a scheduling request.

• PUCCH format 2, capable of transmitting more than two bits and spanning one or two OFDM symbols. This format can, for example, be used for CSI reports or for multi-bit hybrid-ARQ acknowledgments in the case of carrier aggregation or per-CBG retransmission. Three of the formats, 1, 3, and 4, are sometimes referred to as long PUCCH formats as they occupy from 4 to 14 OFDM symbols. The reason for having a longer time duration than the previous two formats is coverage. If a duration of one or two OFDM symbols does not provide sufficient received energy for reliable reception, a longer time duration is necessary and one of the long PUCCH formats can be used.

The long PUCCH formats include:

• PUCCH format 1, capable of transmitting at most two bits.

• PUCCH formats 3 and 4, both capable of transmitting more than two bits but differing in the multiplexing capacity, that is, how many devices that can use the same timefrequency resource simultaneously. Since the PUSH uplink can be configured to use either OFDM or DFT-spread OFDM, one natural thought would be to adopt a similar approach for the PUCCH. However, to reduce the number of options to specify, this is not the case. Instead, the PUCCH formats are in general designed for low cubic metric, PUCCH format 2 being the exception and using pure OFDM only. Another choice made to simplify the overall design was to only support specification transparent transmit diversity schemes. In other words, there is only a single antenna port specified for the PUCCH and if the device is equipped with multiple transmit antennas it is up to the device implementation how to exploit these antennas, for example by using some form of delay diversity. In the following, the detailed structure of each of these PUCCH formats will be described.

PUCCH Format 1:

PUCCH format 1 is to some extent the long PUCCH counterpart of format 0. It is capable of transmitting up to two bits, using from 4 to 14 OFDM symbols, each one resource block wide in frequency. The OFDM symbols used are split between symbols for control information and symbols for reference signals to enable coherent reception. The number of symbols used for control information and reference signal, respectively, is a trade-off between channel-estimation accuracy and energy in the information part. Approximately half the symbols for reference symbols were found to be a good compromise for the payloads supported by PUCCH format 2. The one or two information bits to be transmitted are BPSK or QPSK modulated, respectively, and multiplied by the same type of length-12 low-PAPR sequence as used for PUCCH format 0. Similar to format 0, sequence and cyclic shift hopping can be used to randomize interference. The resulting modulated length-12 sequence is block-wise spread with an orthogonal DFT code of the same length as the number of symbols used for the control information.

The use of the orthogonal code in the time domain increases the multiplexing capacity as multiple devices having the same base sequence and phase rotation still can be separated using different orthogonal codes. The reference signals are inserted using the same structure, that is, an unmodulated length-12 sequence is block-spread with an orthogonal sequence and mapped to the OFDM symbols used for PUCCH reference-signal transmission. Thus, the length of the orthogonal code, together with the number of cyclic shifts, determines the number of devices that can transmit PUCCH format 1 on the same resource. An example where nine OFDM symbols are used for PUCCH transmission, four carrying the information and five used for reference signals. Hence, up to four devices, determined by the shorter of the codes for the information part, can share the same cyclic shift of the base sequence, and a set of resources for PUCCH transmission in this particular example. Assuming a cell-specific base sequence and six out of the 12 cyclic shifts being useful from a delay-spread perspective, this results in a multiplexing capacity of at most 24 devices on the same time frequency resources.

PUCCH Format 2:

PUCCH format 2 is a short PUCCH format based on OFDM and used for transmission of more than two bits, for example, simultaneous CSI reports and hybrid- ARQ acknowledgments, or a larger number of hybrid-ARQ acknowledgments. A scheduling request can also be included in the bits jointly encoded. If the bits to be encoded are too large, the CSI report is dropped to preserve the hybrid- ARQ acknowledgments which are more important. The overall transmission structure is straightforward. For larger payload sizes, a CRC is added. The control information (after CRC attachment) to be transmitted is coded, using ReedMuller codes for payloads up to and including 11 bits and Polar4 coding for larger payloads, followed by scrambling and QPSK modulation. The scrambling sequence is based on the device identity (the C-RNTI) together with the physical-layer cell identity (or a configurable virtual cell identity), ensuring interference randomization across cells and devices using the same set of time frequency resources. The QPSK symbols are then mapped to subcarriers across multiple

resource blocks using one or two OFDM symbols. A pseudo-random QPSK sequence, mapped to every third subcarrier in each OFDM symbol, is used as a demodulation reference signal to facilitate coherent reception at the base station. The number of resource blocks used by PUCCH format 2 is determined by the payload size and a configurable maximum code rate. The number of resource blocks is thus smaller if the payload size is smaller, keeping the effective code rate roughly constant. The number of resource blocks used is upper bounded by a configurable limit. PUCCH format 2 is typically transmitted at the end of a slot .However, similarly to format 0 and for the same reasons, it is possible to transmit PUCCH format 2 also in other positions within a slot.

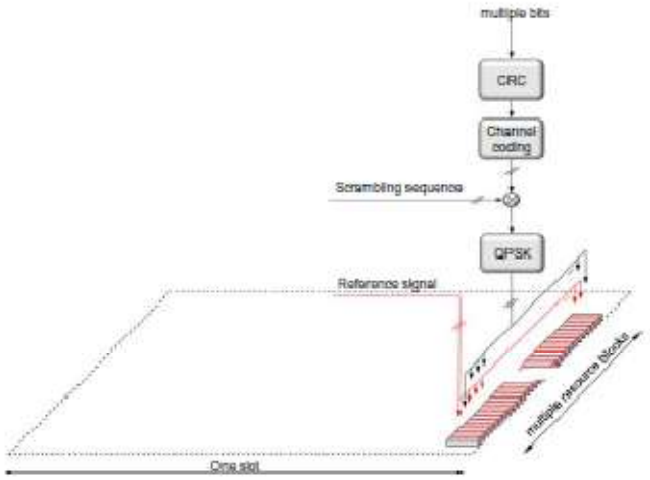


Figure 4.9 PUCCH Format 2

PUCCH Format 3:

PUCCH format 3 can be seen as the long PUCCH counterpart to PUCCH format 2. More than two bits can be transmitted using PUCCH format 3 using from 4 to 14 symbols, each of which can be multiple resource blocks wide. Thus, it is the PUCCH format with the largest payload capacity. Similar to PUCCH format 1, the OFDM symbols used are split between symbols for control information and symbols for reference signals to allow for a low cubic metric of the resulting waveform. The control information to be transmitted is coded using ReedMuller codes for 11 bits or less and Polar codes for large payloads, followed by scrambling and modulation. The scrambling sequence is based on the device identity (the CRNTI) together with the physical-layer cell identity (or a configurable virtual cell identity), ensuring interference randomization across cells and devices using the same set of time frequency resources. Following the principles of PUCCH format 2, a CRC is attached to the control information for the larger payloads. The modulation scheme used is QPSK but it is possible to optionally configure $\pi/2$- BPSK to lower the cubic metric at a loss in link performance. The resulting modulation symbols are divided between the OFDM symbols. DFT precoding is applied to reduce the cubic metric and improve the power amplifier efficiency. The reference signal sequence is generated in the same way as for DFT-precoded PUSCH transmissions (see

Section 9.11.2) for the same reason, namely to maintain a low cubic metric. Frequency hopping can be configured for PUCCH format 3, for example, to exploit frequency diversity, but it is also possible to operate without frequency hopping. The placements of the reference signal symbols depend on whether the frequency hopping is used or not and the length of the PUCCH transmission, as there must be at least one reference signal per hop. There is also a possibility to configure additional reference signal locations for the longer PUCCH durations to get two reference signal instances per hop. The mapping of the UCI is such that the more critical bits, that is, hybrid- ARQ acknowledgments, scheduling request, and CSI part 1, are jointly coded and mapped close to the DM-RS locations, while the less critical bits are mapped in the remaining positions.

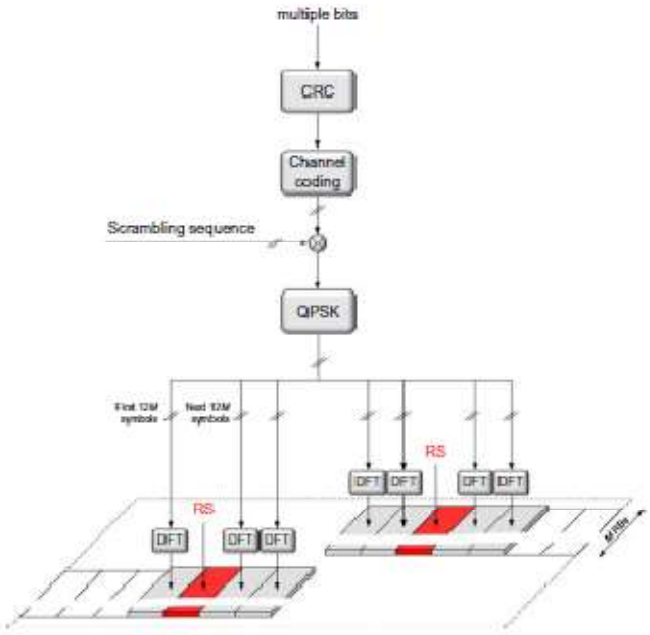


Figure 4.10 PUCCH Format 3

PUCCH Format 4:

PUCCH format 4 is in essence the same as PUCCH format 3 but with the possibility to code-multiplex multiple devices in the same resource and using at most one resource block in the frequency domain. Each controlinformation- carrying OFDM symbol carries 12/NSF unique modulation symbols. Prior to DFT-precoding, each modulation symbol is block-spread with an orthogonal sequence of length NSF. on the same set of resource blocks.

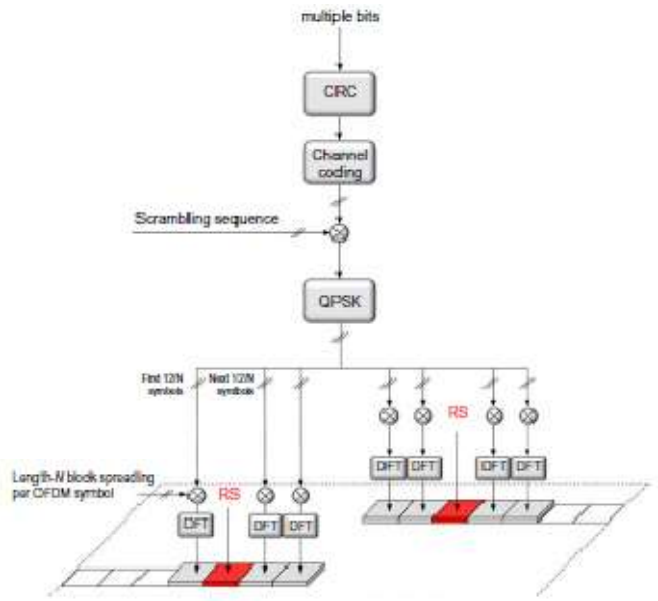


Figure 4.11 PUCCH Format 4

Spreading factors two and four are supported, implying a multiplexing capacity of two or four devices.

4.8 RESOURCES AND PARAMETERS FOR PUCCH TRANSMISSION

In the discussion of the different PUCCH formats above, a number of parameters were assumed to be known. For example, the resource blocks to map the transmitted signal to, the initial phase rotation for PUCCH format 0, whether to use frequency hopping or not, and the length in OFDM symbols for the PUCCH transmission. Furthermore, the device also needs to know which of the PUCCH formats to use, and which time frequency resources to use. In LTE, especially in the first releases, there is a fairly fixed linkage between the uplink control information, the PUCCH format, and the transmission parameters. For example, LTE PUCCH format 1a/1b is used for hybrid-ARQ acknowledgments and the time frequency-code resources to use are given by a fixed time offset from the reception of the downlink scheduling assignment and the resources used for the downlink assignment. This is a low-overhead solution, but has the drawback of being inflexible and was extended to provide more flexibility in later releases of LTE supporting carrier aggregation and other more advanced features. NR has adopted a more flexible scheme from the beginning, which is necessary given the very flexible framework with a wide range of service requirements in terms of latency and spectral efficiency, support of no predefined uplink downlink allocation in TDD, different devices supporting aggregation of different number of carriers, and different antenna schemes requiring different amounts of feedback just to name some motivations. Central in this scheme is the notion of PUCCH resource sets. A PUCCH resource set contains at least four PUCCH resource

configurations, where each resource configuration contains the PUCCH format to use and all the parameters necessary for that format. Up to four PUCCH resource sets can be configured, each of them corresponding to a certain range of UCI feedback to transmit. PUCCH resource set 0 can handle UCI payloads up to two bits and hence only contain PUCCH formats 0 and 1, while the remaining PUCCH resource sets may contain any PUCCH format except format 0 and 1. When the device is about to transmit UCI, the UCI payload determines the PUCCH resource set and the ARI in the DCI determines the PUCCH resource configuration within the PUCCH resource set. Thus, the scheduler has control of where the uplink control information is transmitted. For periodic CSI reports and scheduling request opportunities, which both are semistatically configured, the PUCCH resources are provided as part of the CSI or SR configuration.

Reference Books:

1. Saad Z. Asif, "5G Mobile Communications Concepts and Technologies, CRC Press, 1st Edition, 2019.
2. Erik Dahlman, Stefan Parkvall, Johan Skold "5G NR: The Next Generation Wireless Access Technology", Academic Press, 1st Edition, 2018.
3. Jonathan Rodriguez, "Fundamentals 5G Mobile Networks", John Wiley & Sons, 1st Edition, 2015.
4. Long Zhao, Hui Zhao, Kan Zheng, Wei Xiang, "Massive MIMO in 5G Networks: Selected Applications", Springer, 1st Edition, 2018.
5. Robert W. Heath Jr., Angel Lozano, "Foundations of MIMO Communication", Cambridge University Press, 1st Edition, 2019.
6. R. Vannithamby and S. Talwar, "Towards 5G: Applications, Requirements and Candidate Technologies", John Willey & Sons, 1st Edition, 2017.

Questions to Practice:

PART -A

1	Organise about Report Quantity in 5G NR
2	Examine in detail about Multiport SRS
3	Support your answer how Channel Coding plays a vital role Physical Channel Processing
4	Assess 5G Core Network Scrambling in detail
5	Classify Resource Mapping Tlechniques in detail

PART-B

1	Support the statement how Downlink Measurements and Reporting is essential
2	Justify Transport Channel Processing in 5G Technology
3	Precoding is a generalization of beam forming to support multi-stream (or multi-layer) transmission in multi-antenna wireless communications. Support the above statement that relates to Multi Antenna Precoding
4	Investigate in detail about Physical Downlink Control Channel in 5G

Communication.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF ELECTRICAL AND ELECTRONICS

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

UNIT – V –ENABLING TECHNOLOGIES FOR 5G– SECA3020

UNIT-V (ENABLING TECHNOLOGIES FOR 5G)

Device-to-Device (D2D) Communication - 5G for Massive Machine Type Communication and Massive IoT- V2X Communication - Full Duplex and Green Communication - mmWave Communications -Massive MIMO and Beamforming Techniques

5.1 Introduction :

Massive Multiple-Input and Multiple-Output (MIMO) technology is an important and timely topic, which is largely motivated by the requirements of the Fifth Generation (5G) or future wireless communications. By offering a large number of Degrees of Freedom (DoF), 5G is capable of simultaneously serving multiple users with high gains and thus improving the system Spectrum Efficiency (SE), Energy Efficiency (EE) and reliability. Different from existing studies in the literature, this book focuses specifically on the state-of-the-art of massive MIMO and its typical applications, such as millimeter-wave (mm-wave) communications and wireless energy transfer. In this chapter, we first present the motivations of massive MIMO following a short overview on the requirements and techniques of 5G communications. Then, basic concepts alongside the pros and cons of both MIMO and massive MIMO systems are given.

Mobile Internet and the Internet of Things (IoT) are two key market drivers for 5G communications. The applications of 5G communications include cloud computing, eHealth services, automotive driving, tactile Internet, Augmented Reality (AR)/Virtual Reality (VR), Cyber-Physical System (CPS) and so on. These applications could be classified into three use scenarios, i.e., Enhanced Mobile Broadband (eMBB), Massive Machine Type Communications (mMTC), and Ultra- Reliable and Low Latency Communications (URLLC). In order to support the requirements of the three use scenarios, some performance targets of 5G communications are defined, i.e., increasing throughput by 1000- fold, improving EE by 10-fold, shortening the end-to-end delay to 1/5–1/10, increasing the number of connected equipment by 10–100 folds, and prolonging the battery lifespan of low-power equipment by 10-fold. Some research projects for 5G communications have been launched at home and abroad in order to achieve the performance targets of 5G before 2020 year.

There are three key techniques for 5G communications from the perspective of system capacity. That is, the massive MIMO technique is first adopted to improve system SE; mm-wave spectral resources are employed to expand system bandwidth; and multi-layer and ultra-dense networks are deployed to increase geographic spectral reuse. The systems employing massive antenna arrays to serve multiple users are dubbed massive MIMO communication systems. Massive MIMO systems are capable of combatting the severe fading of mm-wave signals, providing the wireless backhaul, and suppressing interference in multi-layer and denser networks. Therefore, this monograph focuses on massive MIMO technology and their typical application scenarios in an attempt to provide the basic theory and paradigms for practical system designs.

5.2 MIMO TECHNOLOGY

Bandwidth Efficiency (BE) or SE is usually one of the most important metrics to select candidate technologies for next-generation wireless communications systems. Meanwhile, with excessive power consumption in wireless communications networks, both carbon emissions and operator expenditure increase year by year. As a result, EE has become another significant metric for evaluating the performances of wireless communications systems with some given BE constraints.

5.2.1 Traditional MIMO

MIMO technology has attracted much attention in wireless communications, because it offers significant increases in data throughput and link range without an additional increase in bandwidth or transmit power. In 1993 and 1994, a MIMO approach was proposed and the corresponding patent was issued, where multiple transmit antennas are co-located at one transmitter with the objective of improving the attainable link throughput. Then, the first laboratory prototype of spatial multiplexing was implemented to demonstrate the practical feasibility of MIMO technology. Nowadays, MIMO has been accepted as one of key technologies in Fourth Generation (4G) wireless communication systems. When an evolved Node B (eNB) equipped with multiple antennas communicates with several User Equipments (UEs) at the same time-frequency resources, it is referred to as Multi-User MIMO (MU-MIMO). MU-MIMO is capable of improving either the BE or the reliability by improving either the multiplexing gains or diversity gains

5.2.2 Massive MIMO

In order to scale up these gains of traditional MIMO, the massive MIMO concept, which is also known as Large-Scale MIMO (LS-MIMO) scheme often also associated with the terminologies of large-scale antenna systems, very large MIMO, very large MU-MIMO, full-dimensional MIMO, hyper MIMO, etc., was proposed by Marzetta. More explicitly, massive MIMO refers to the system that uses hundreds of antennas to simultaneously serve dozens of UEs. Both theoretical and measurement results indicate that massive MIMO is capable of significantly improving the BE, which simultaneously reducing the transmit power. As a result, massive MIMO is regarded as a candidate technique for next-generation wireless communications systems conceived for the sake of improving both their BE and EE. As the down tilt of an Antenna Array (AA) is fixed, traditional MIMO technology can only adjust signal transmission in the horizontal dimension. In order to exploit the vertical dimension of signal propagation, AAs, such as rectangular, spherical and cylindrical AAs, were studied by the 3rd Generation Partnership Project (3GPP). MIMO with these arrays can adjust both azimuth and elevation angles, and propagate signals in Three-Dimensional (3D) space, thus termed 3D MIMO. To further increase capacity, 3D MIMO deploys more antennas to achieve larger multiplexing gains. Meanwhile, massive MIMO adopts rectangular, spherical or cylindrical AAs in practical systems considering the space of AAs. Therefore, 3D MIMO with massive antennas can be seen as a practical deployment means of massive MIMO. Massive MIMO can improve BE since it can achieve large multiplexing gains when serving tens of UEs simultaneously. The

significant increase in EE is due to the fact that the use of more antennas helps focus energy with an extremely narrow beam on small regions where the UEs are located. Apart from these advantages, massive MIMO can enhance transmission reliability owing to the excessive DoF. Inter-User Interference (IUI) can also be alleviated because of the extreme narrow beam. In a massive MIMO system, individual element failure of the AA is not detrimental to the performance of the entire system. Simple low-complexity signal processing algorithms are capable of approximating the performance achieved by optimal methods, such as Maximum-Likelihood (ML) multiuser detection and Dirty Paper Coding (DPC). The latency of the air interface can be reduced and the protocols at the Media Access Control (MAC) layer can be simplified because of the channel harden phenomenon and sufficient capacity Certainly, the complexity of signal processing, including Transmit Precoding (TPC), channel estimation and detection, increases with the increasing number of antennas. On the other hand, the maximum number of orthogonal pilot sequences is limited by the coherence interval and coherence bandwidth. Therefore, the performance of massive MIMO systems is constrained by pilot contamination due to pilot reuse in multi-cell scenarios. Moreover, compared to the Physical Downlink Shared Channel (PDSCH) employing either precoding or beamforming, the Signal-to-Interference-plus-Noise Ratio (SINR) of the Physical Broadcasting Channel (PBCH) is lower due to the omni-directional signal transmission. According to the current literature related to massive MIMO, the major research directions about massive MIMO are listed in Table 1.1, some of which have been investigated and while others are not.

5.3 HOMOGENEOUS NETWORK SCENARIOS

5.3.1 Multi-Layer Sectorization

Upon increasing the number of UEs and their carried tele-traffic in urban environments, increased system capacity is required for supporting customer requirements. Traditionally, sectorization techniques are used for providing services to a growing population, which simply divide a cell into multiple sectors, thus increasing network capacity. The equipment costs can also be reduced by allowing a single eNB to serve either three 120o sectors or six 60o sectors. However, although sectorization is capable of improving the area BE, this benefit comes at the expense of a potentially increased interference among sectors due to non-ideal sector-antenna patterns. Therefore, more efficient techniques are required to further increase the achievable network capacity. As illustrated in Fig. 5.1, accurate sectorization in massive MIMO systems can be achieved by high-selectivity angular beamforming performed horizontally, which is capable of reducing the interference among sectors. Moreover, the coverage of each beam can be changed by adjusting the elevation angle of 3D beamforming. By this way, a conventional fixed sector can be further spitted into inner and outer sectors, each of which can be served by a 3D Beamformer (BF) with the same horizontal but different elevation angles. The same frequency radio resources are reused by all the sectors, which is capable of significantly increasing the number of UEs served and/ or of improving the network’s throughput.

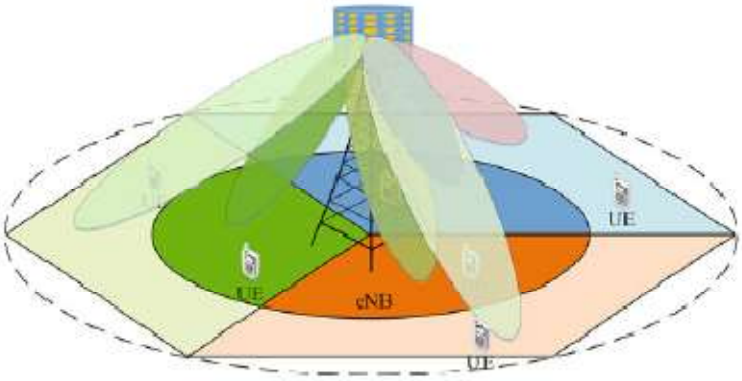


Figure 5.1 Multilayer Sectorization

5.3.2 Adaptive Beamforming

Fixed BFs are so called because the weights that multiply the signals at each element of the AA remain unchanged during operation. By contrast, the weights of an adaptive BF are continuously updated based on the received signals in order to suppress spatial interference, e.g., as depicted in Fig. 5.2. This process may be carried out in either the Time-Domain (TD) or Frequency Domain (FD). Compared to the Two-Dimensional (2D) adaptive BF, a 3D BF may have more flexibility in reusing the radio resources in the spatial domain.

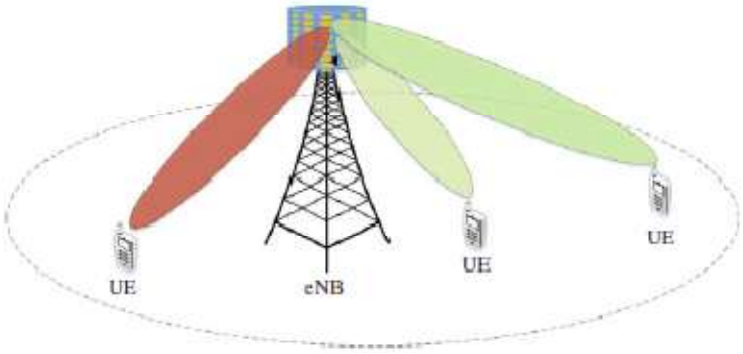


Figure 5.2 Adaptive Beamforming

5.3.3 Large Scale Corporation

Most of the existing contributions on massive MIMO show different benefits in a co-located deployment scenario, where there is a large number of antennas installed at a single cell cite. However, such co-located deployments impose challenges both on their hardware design and on their field deployment. On the other hand, Distributed Antenna Systems (DASs) associated with spatially separated antennas have been conceived for improving the indoor coverage using a moderate number of antennas. Recent studies have shown that apart from its improved coverage, a DAS is capable of significantly increasing the network's BE, even in the presence of Inter-cell Interference (ICI). This motivates researchers to identify specific scenarios as illustrated in Fig. 5.3, where the massive MIMO system associated with a distributed architecture outperforms the one relying on a co-located deployment.

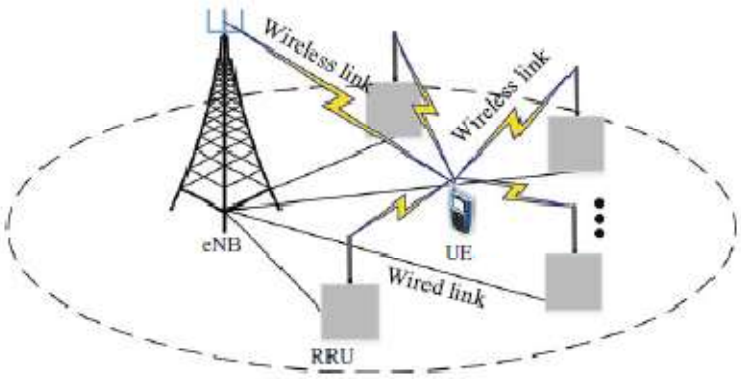


Figure 5.3 Large Scale Cooperation

The advantage of distributed massive MIMO is plausible, because the signals arriving from the distributed antennas to each UE are subject to independent random levels of large-scale fading, thereby leading to potential capacity gains over their colocated counterpart. However, it may be a challenge to achieve these gains by coordinating the intra-cell interferences, especially in scenarios having dozens or even hundreds of Remote Radio Units (RRUs) in a cell. Although full cooperation constitutes an efficient method of eliminating the intra-cell interference, it is not practical due to its high reliance on full Channel State Information (CSI) sharing. To strike an elegant trade-off between the performance attained and the overhead imposed, efficient large-scale cooperation schemes are of high importance under this scenario. Moreover, distributed massive MIMO and small cell deployments may be viewed as being complementary rather than competitive. For example, a cooperative cellular architecture composed of a DAS and a femto cell-macro cell underlay system, which may be extended to operate in conjunction with distributed massive MIMO.

5.4 HETEROGENEOUS NETWORK SCENARIOS

5.4.1 Wireless Backhaul:

The HetNet with dense small cells has been regarded as a very promising design architecture in terms of energy and area BE. It typically consists of multiple types

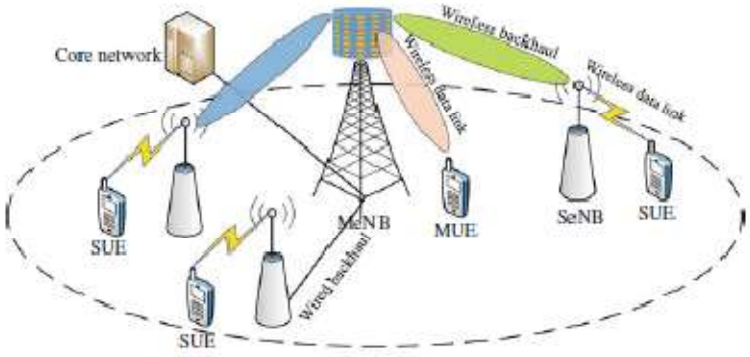


Figure 5.4 Wireless Backhaul

of radio access nodes, e.g., aMacro-cell eNB (MeNB) and multiple Small-cell eNBs (SeNBs) such as pico, femto and relay eNBs. All SeNBs need to be connected to their donor MeNBs through a wired or wireless backhaul. Generally, the wireless backhaul is preferred to instead of the wired backhaul because of easy deployment. In this scenario, a massive MIMO is used at the MeNB, which has a high DoF so to support multiple wireless backhauls in the HetNet. The same spectrum may be reused among wireless backhauls, access of Macro-cell UEs (MUEs) and Small-cell UEs (SUEs). In other words, SeNBs can be viewed as a special kind of UEs communicating with the MeNB via the wireless backhaul. Since the location of an eNB is usually fixed, the channel of the wireless backhaul may be quasi-static time varying. Therefore, the MeNB is capable of eliminating the interference between the wireless backhaul and MUEs through the use of precoding.

5.4.2 Hotspot Coverage

Statistics show that the majority of tele-traffic originates from buildings, such as super-markets, office buildings, gymnasiums. Therefore, high quality indoor coverage of buildings is considered as one of the key scenarios for the HetNet. Since the tele-traffic is generated at different heights in buildings, traditional AAs with a fixed Downlink (DL) tilt, which are mainly designed for UEs roaming at the street level, are no longer suitable for this scenario. A massive AA is capable of dynamically adjusting both the azimuth and elevation angles of its beam. It can transmit the beams directly to the UEs at different floors in a building, and thus significantly improves system throughput.

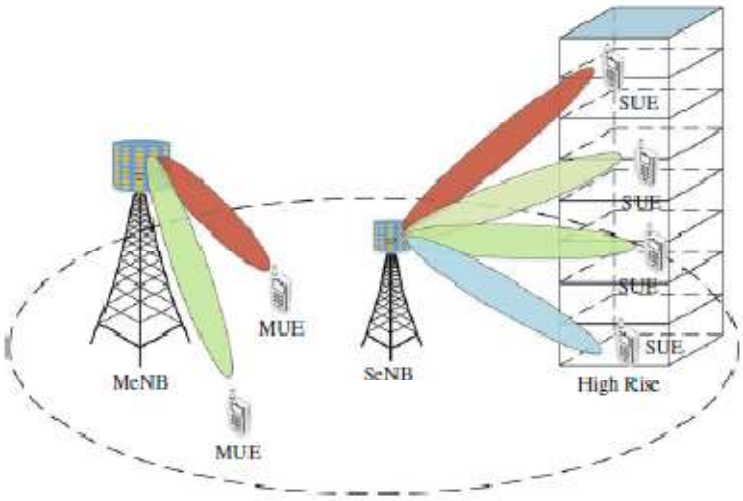


Figure 5.5 Hotspot Coverage

However, when the indoor coverage of the building is provided by the MeNB with a massive AA, the adjustable range of the elevation angles remains small compared to that of the SeNB, and the angular resolution cannot meet the needs of UEs. As it is well known, the close distance between SeNBs and SUEs results in reduced path losses. Therefore, the SeNBs equipped with a massive AA are more appropriate for in-building coverage, providing that deployment costs are acceptable.

5.4.3 Dynamic Cell:

Since the Reference Signal Received Power (RSRP) gleaned from the MeNB is usually higher than that from the SeNB in HetNets, more UEs are likely to be connected to the MeNB, leading to a potential unbalanced traffic distribution between the macrocell and small cells. The Cell Range Extension (CRE) technique may be used for offloading the traffic from the macro-cells to small cells [9]. However, the UEs in the extended range, which are somehow forced to access to the small cells, may experience low SINRs due to the strong interference encountering from the MeNB. This may cause the unreliable communications between them and SeNBs.

In order to solve this problem, the Almost Blank Subframe (ABS) technique can be applied to reduce the interference from the MeNB through time domain coordination. In other words, the Quality of Service (QoS) performances of SUEs in the extended range is improved at the expense of multiplexing gains. With the introduction of massive AAs into SeNBs, the down tilt of the transmit signals is adjustable achieve a better received signal quality at SUEs. As illustrated in Fig. 5.6, it is helpful in adaptively expanding or shrinking the radius of small cells, i.e., Dynamic cell. Therefore, the UEs at the edge of the small cell may opt for adaptively connecting to the SeNB according to their received power level. It is appropriate for balancing the traffic between the macro-cell and small cells in the extended range.

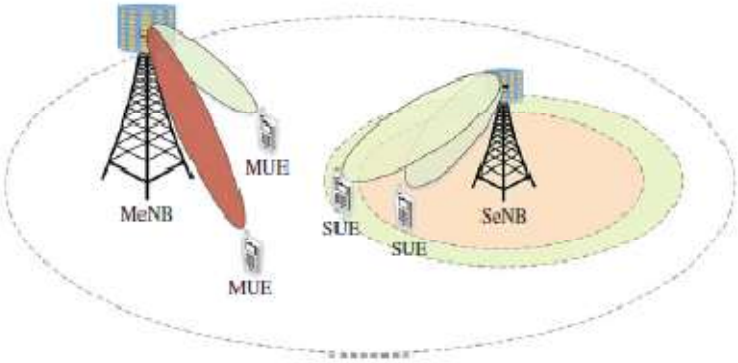


Figure 5.6 Dynamic Cell

In this section, typical application scenarios of massive MIMO are classified into two types, i.e., homogeneous and heterogeneous networks with massive MIMO. The former is with only macro-cell deployment, which includes multi-layer sectorization, adaptive beamforming and large-scale cooperation. Multi-layer sectorization is capable of increasing multiplexing gains through splitting the sectors. Adaptive beamforming focuses the radiated energy in the desired direction with the aid of an extremely narrow beam, which is able to improve the desired SINR of the UE, whilst simultaneously reducing the interference imposed on other UEs. Compared to the conventional DAS technique, large-scale cooperation, through scaling up the number of distributed antennas that are coordinated, is capable of further enhancing both coverage and achievable throughput.

There are three typical application scenarios in the case of the HomoNet with massive MIMO. Firstly, the employment of a wireless backhaul by massive MIMO between MeNB and SeNBs is more flexible and of low cost compared to its wired backhaul counterpart. Then, the SeNB with a

massive AA is able to adaptively adjust both the azimuth and elevation angles in an effort to improve the coverage of indoor hotspots, e.g., in buildings.

Moreover, the cell radius in HetNet is dynamically adjustable so as balance the load between MeNB and SeNBs by changing the elevation angle. Based upon the above discussions, massive MIMO is expected to be applied in numerous scenarios to improve achievable capacity and throughput. However, extensive studies are still needed in practical network deployment.

5.5 NETWORK TECHNOLOGY

While the underlying physical layer techniques lay the foundation of massive MIMO systems, networking techniques also play a vital role in practical systems, making them operate more efficiently, reliably and securely. Due to their crucial impact on the attainable performance of massive MIMO systems, networking techniques have gradually attracted considerable interest in both academia and industry. The effective exploitation of radio resources is one of the main goals to be achieved by networking techniques.

Towards this end, several performance indicators are considered, such as the aforementioned BE and EE. Always serving UEs which are experiencing the best channel conditions is surely capable of improving the system's BE. However, this may result in unfair resource allocation, potentially disadvantaging those UEs that suffer from poor channel conditions, such as the UEs located at cell edges. Therefore, apart from BE and EE, the networking techniques usually take into account fairness in order to guarantee a certain level of minimum performance for all UEs. In the remainder of this section, we only focus our attention on two networking techniques, i.e., Inter-cell Interference Coordination (ICIC) and radio resource scheduling, which are two most important issues in wireless networks.

5.5.1 Inter-Cell Interference Coordination

Cellular communication systems suffer from ICI at the cell boundaries, especially when all the channels are fully reused in adjacent cells. As a result, interference mitigation and coordination techniques are needed for alleviating ICI so as to well support frequency reuse. Here, we focus only on static or semi-static ICIC approaches for massive MIMO systems in different network deployments.

Homogeneous Networks:

ICIC techniques, such as Fractional Frequency Reuse (FFR) and Soft Frequency Reuse (SFR), have been widely investigated in the context of efficient radio resource management in multi-cell environments in an attempt to coordinate co-channel interference, resulting in improved cell-edge coverage, cell edge data rates and area BE. A large-scale AA provides additional spatial DoF. Therefore, ICIC for massive MIMO systems is able to exploit the spatial DoF for mitigating ICI by nulling certain spatial direction to the neighbouring cell. In a massive MIMO system, each eNB is equipped with a huge number of antennas, serving its scheduled UEs with beamforming, while trading off its excess DoF against coordinating the interference to other cells within a cluster.

Compared with network MIMO, massive MIMO is preferred to because of its low costs of deploying an excessive number of antennas at the cell site. Under the assumption of the same number of DoF per UE and same amount of channel estimation overhead, massive MIMO with spatial interference coordination outperforms network MIMO. The 3D MIMO system, one of massive MIMO systems, has the capability to dynamically adapt the shape of the vertical beamforming pattern to the UEs at different locations. In other words, the UEs at the cell center and cell edge are covered by different vertical beamforming patterns with specific downtilt such that the received signal power for each UE is maximized.

Then, cell sectorization in the 3D MIMO system can be carried out not only along the horizontal but also the vertical axis, which results in increased system throughput. However, the ICI problem becomes more complicated with much more sectors per cell. Therefore, it is not straightforward whether the overall BE performance as well as the cell edge UE can be improved. The preliminary study in [58] shows that dynamic vertical beam pattern adaptation can provide BE performance gains even with either simplified or suboptimum approaches. Meanwhile, there exists some work on coordinated vertical beamforming with well-known ICIC schemes applied in Long Term Evolution (LTE) such as FFR. In the literature, there is a lack of comprehensive studies on this issue to date.

Heterogeneous Networks

The HetNet is an attractive means of increasing achievable network capacity and of enhancing the coverage area and/or Quality of Experience (QoE). In a HetNet, small cells as a tier are capable of providing hotspot capacity enhancements, whereas macro cells as another tier are responsible for large area coverage in support of high mobility UEs. However, the MeNBs and SeNBs may interfere with each other, if they use the same time-frequency resources without careful coordination. Fortunately, when the MeNBs, or even the SeNBs, are equipped with a large-scale AA, the AA can provide an additional spatial DoF for multiplexing the data of several UEs onto the same time-frequency resource.

Furthermore, it can concentrate the radiated energy precisely on the intended UEs, thereby reducing both the intraand inter-tier interference. Massive MIMO systems are also capable of supporting cooperation in an implicit way between the different tiers in the HetNet for the sake of improving the overall system performance. To satisfy ever increasing data rate demands, a two-tier TDD-based HetNet is introduced, where the macro-cell tier served by the MeNBs equipped with a large-scale AA is overlaid with the small cell tier of single-antenna SeNBs. Making use of explicit benefits of channel reciprocity under the TDD mode, the MeNBs estimate the UL interference covariance matrix characterizing the interference from the overlay small cells, which can be used for DL ZF based TPC to reduce the interference to the SUEs. The MeNBs with massive MIMO can significantly improve the BE of small cells at the expense of a moderate loss of the macro cell performance. Additionally, the SeNBs can also be equipped with multiple antennas if needed.

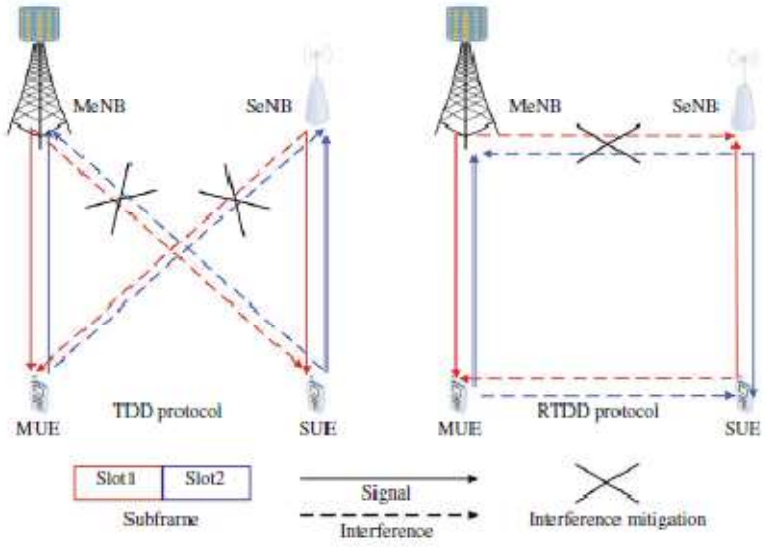


Figure 5.7 TDD or RTDD

Recently, the so-called Reversed TDD (RTDD) protocol has been proposed for the HetNet [62]. In the RTDD protocol as shown in Fig. 5.7, the sequence of the UL and DL transmission periods in one of the tiers is reversed to the other. For example, in Slot 1, while the MeNB transmits the signals to MUEs in the DL, the SeNB received the signals from SUEs in the UL, and vice versa. In the traditional TDD protocol, the MeNBs and SUEs interfere with each other, and so do the SeNBs and MUEs. The channels between the eNBs and UEs potentially fluctuate rapidly thanks to UE mobility. Therefore, less interference samples are available for approximating the time-averaged interference covariance, and hence the resultant estimation errors may degrade the attainable system performance.

However, the interference scenario in the HetNet is different if the RTDD protocol is applied, where the MeNB and SeNB interfere with each other, and so do the MUEs and SUEs. Since both the MeNB and SeNB are fixed in location, the interference between them are quasi-static. Hence, the estimated covariance of the channels between them are not sensitive to instantaneous channel variation. Moreover, a massive AA can be deployed at the MeNB and even at the SeNBs so that the interference between the eNBs can be nearly eliminated by narrow beamforming. Meanwhile, the interference between UEs is usually not very serious in most cases because of the low transmission power of UEs. As a result, the RTDD protocol is more suitable for the HetNet with massive MIMO, since it readily lends itself to cooperative interference cancellation.

5.5.2 Scheduling

Based on the status of queue, channel quality, QoS requirements and so on, the eNB schedules limited radio resources across the time, frequency and spatial domains among the UEs. Different design objectives, e.g., the affordable complexity, overhead, BE, and fairness, are targeted by a variety of scheduling schemes. Usually a good trade-off among all these goals is strived for practical wireless communication systems. Given a large number of UEs and a limited number of antennas, the problem of sum capacity scaling with UE selection has been widely investigated. In

particular, opportunistic beamforming yields significant gains by exploiting the independence of the UEs' channel fluctuation, which may be conducive to achieving multiple UE diversity.

Moreover, in order to evaluate the gains of scheduling, the mutual information of the massive MIMO system may be modelled as a normal distribution under the assumption of i.i.d. Rayleigh fading channels. It turns out that the variance of the channel coefficients grows slowly or even decreases with the number of transmit antennas, but increases with the number of UEs. This conclusion implies that carefully designed robust scheduling schemes would reduce the CSI-feedback rate required. The more UE-specific and eNB-specific channel information is obtained by the scheduler, the more efficient the system can be. Nonetheless, with increased amount of required information, the overhead and computational complexity for channel information may become prohibitive. Due to the channel hardening phenomenon of massive MIMO systems, the scheduling mechanisms relying on full CSI, including large-scale and small-scale prorogation characteristics, are not costefficient, because no significant performance gains can be achieved compared to those requiring only partial CSI including the path loss and shadow fading. So, more attention has been paid to scheduling schemes with partial CSI instead of full CSI in massive MIMO systems, both of which are presented for comparison in this section.

Full CSI-Based Scheduling

In order to achieve the optimal network performance, the scheduler has to acquire the full and accurate knowledge of all the channels. This knowledge can be exploited to minimize the total power consumption, while satisfying the QoS and power constraints at the eNBs in the HetNet with massive MIMO. Toward this end, a spatial soft-cell approach can be taken, where each UE is dynamically assigned to access the optimal nodes, i.e., massive MIMO MeNB, conventional SeNB or both. If the system assigns a UE to an MeNB and an SeNB at the same time, multiple transmitters serve the UE through joint non-coherent multi-flow beamforming. Also, the total EE can be further improved by applying a low-complexity efficient algorithm, which exploits the hidden convexity in this problem. However, the scheduler has to know the full instantaneous CSI of all the UEs, which incurs both potentially overwhelming estimation overheads and excessive computational complexity due to the large number of AEs and UEs in a practical system. The costs of the scheduling scheme based on full CSI are likely to outweigh the gains.

Partial CSI-Based Scheduling

Fortunately, scheduling schemes with partial CSI can reduce complexity and overhead of massive MIMO systems with an acceptable performance loss.

Single-Cell Scheduling: Under the single-cell scenarios, resource allocation for energy-efficient Orthogonal Frequency Division Multiple Access (OFDMA) systems with a large number of AEs. Taking into account the associated circuit power dissipation, the imperfect CSI at the Transmitter (CSIT) and QoS requirements, the resources are assigned with the objective of maximizing EE. The considered parameters include the subcarrier allocation, power allocation, antenna allocation and available data rates. It is demonstrated that even though the use of a large number of transmit antennas reduces the multipath-induced fluctuation of each channel, the system performance can still benefit from the different path losses and shadow fading conditions

of different UEs. Furthermore, the scheduling scheme can be updated periodically, because the path loss and shadow fading parameters vary slowly, depending on UE mobility. The reduced-complexity probabilistic scheduling algorithms have been proposed, where the number of AEs is quite large. Firstly, different UEs are clustered into groups based on their channel covariance matrices. Then, the UEs in each group are pre-selected randomly based on the group-specific probabilities derived from the whole system. Only the pre-selected UEs are required to transmit their training signals or CSI feedback, leading to markedly reduced overheads and complexity. Moreover, different measures have been conceived both for UE grouping and for scheduling in a FDD based massive MIMO system based on twostage precoding, namely on inter-group precoding and intra-group precoding, which are capable of reducing the channel estimation overhead while guaranteeing fairness to the UEs.

Multi-Cell Scheduling: Unlike the single-cell scenario, both inter-cell and intracell interferences have to be taken into account under the multi-cell scenario. Although fully multi-cell scheduling schemes, like the coordinated joint processing in network MIMO, can achieve large performance gains, they are not practical due to prohibitive costs. That is, not only the full CSI but also data streams intended for different mobile users at different cells need to be shared among the eNBs. Instead, another feasible solution is to improve the overall network performance by allowing beamforming vectors from different eNBs to be coordinated for the sake of implementation. The main objective of applying coordinated beamforming in massive MIMO is to improve the overall system performance, while reducing the coordination overhead. There are two popular coordinated beamforming schemes, namely the hierarchical and the nested structure . Explicitly, the BF relying on the hierarchical structure at each eNB consists of an inner precoder and an outer precoder. The inner precoder supports the transmission of data to the serving UEs by exploiting the knowledge of the timevariant CSI. Meanwhile, the outer precoder exploits the remaining spatial DoF for mitigating the ICI by relying only on the knowledge of the average CSI. This structure requires only a modest amount of backhaul overhead and pilots for CSI estimation. Moreover, only the knowledge of the channel's spatial correlation matrices is needed for the BF-weight optimization, which is insensitive to the backhaul latency.

By contrast, in the BF associated with the nested structure, the optimal strategy can be found recursively, where the BF weight optimization objective may be based on the fairness in power usage subject to satisfying the target SINR constraints. Another nested structure, whose OF focuses on Maximizing the Minimum (MAX-MIN) weighted SINR among UEs was proposed in . Unlike the hierarchical structure, the optimal precoder of the nested structure is found as the solution of a joint optimization problem, which aims for striking a trade-off between providing a high SINR for the intra-cell UEs and mitigating the ICI. For instance, when this BF is applied in a massive MIMO system, all the eNBs are divided into two groups, (1) Selfish eNBs, whose UE SINRs are relative low; and (2) altruistic eNBs, whose UEs SINRs are relative high. The altruistic group may be empty; or it may use zero-forcing for eliminating the interference by imposing it on the selfish group. If the optimal ZF beamforming scheme is used in the altruistic group, each eNB in this group has to transmit less power than that in the selfish group. The precoder at each eNB uses the optimal BF parameters along with its own instantaneous CSI. Additionally, only the average CSI has to be exchanged among the eNBs. On the other hand,

different metrics can be used to indicate different system performances. Thus, it is crucial to choose an objective with an appropriate metric function for the multi-cell scheduling schemes. There are several kinds of objectives in terms of either efficiency or fairness when multi-cell scheduling is applied, e.g., (a) Minimizing the Maximum (MIN-MAX) the fairness in power consumption subject to certain SINR constraints: It aims to maximize the overall EE as a high priority by adjusting its coordinated beamforming scheme. A efficient solution can be obtained through Lagrange duality and random matrix theory; (b) MAX-MIN SINR subject to certain sum-rate constraints [75, 77]: It enforces the overall system fairness by guaranteeing each UE's promised SINR. In the case of non-convex optimization problems, the optimal scheme may be derived by using nonlinear Perron-Frobenius theory; (c) Maximizing the weighted sum rate subject to some eNBs power consumption constraints: This objective can be viewed as a combination of MAX-MIN fairness and maximum sum-rate. Moreover, efficient schemes can be obtained through hidden convexity and random matrix theory. In this section, a wide range of networking techniques conceived for massive MIMO systems have been investigated, with an emphasis on the associated ICIC and scheduling issues. As for ICIC, the beamforming for massive MIMO systems, which helps eliminate the ICI at the expense of computational, and 3D MIMO beamforming, which can be used for cell splitting with low complexity, have been discussed in the context of the HomoNet. The R-TDD protocol, which is helpful for interference cancellation, has also been studied in the context of the HetNet. In scheduling schemes, existing algorithms have been classified according to their requirements in terms of CSI and their different design objectives. The primary objective of scheduling is to improve the attainable system performance, while maintaining affordable implementation complexity and overhead.

5.6 MASSIVE MIMO AIDED MILLIMETER COMMUNICATION TECHNOLOGY:

Heterogeneous and Small Cell Networks (HetSNets) can increase the SE and throughput with the hierarchical deployment of low power nodes as well as the macro nodes. In order to meet the increasing capacity requirements of future 5G wireless networks, much more number of low power nodes needs to be deployed if no new frequency resources becomes available, which may cause the serious interference between nodes. Then, the introduction of mm-wave communications with massive MIMO provides unprecedented spectral resources for HetSNets in 5G. However, major challenges remain for implementing mm-wave massive MIMO in HetSNets. To this end, we introduce several typical deployment scenarios for HetSNets with mm-wave massive MIMO. A frame structure based on TDD is proposed and discussed then in details. Next, the challenges and possible solutions relating to both physical,MAC and network layers are studied. Finally, system-level simulations are implemented to evaluate the system performance of HetSNets with mm-wave massive MIMO.

Reference Books:

1. Saad Z. Asif, "5G Mobile Communications Concepts and Technologies, CRC Press, 1st Edition, 2019.