J. Indian Inst. Sci.
A Multidisciplinary Reviews Journal
ISSN: 0970-4140 Coden-JIISAD

REVIEW ARTICLE

5G and Beyond: Physical Layer Guiding Principles and Realization

Parvathanathan Subrahmanya[*] *and Amir Farajidana*

Abstract | The term 5G refers to the fifth generation of cellular wireless standards. Following the first four generations which focused on evolving voice and data communications, 5G is expected to greatly expand the range of possible domains to which cellular communications can be applied. The aim of this article is to present an introduction to some guiding principles that provide shape to the physical layer of the 5G standard.

Keywords: *5G, NR, Cellular, Millimeter wave, Low latency*

1 Introduction

Wireless communication has become an integral part of daily life. Over the last four decades, a number of wireless technologies have been developed to serve a variety of use cases. For instance, low-power short-range communication between devices such as smartphones and wireless headsets, or laptops and wireless keyboards up to a few metres apart, can be accomplished with Bluetooth. For higher data rate communication across several tens of metres, Wi-Fi is embedded in a wide variety of devices such as laptops, tablets, TVs, smartphones, etc. The focus of this article is on Cellular Communication—wireless technology that enables communication across hundreds of metres up to several kilometres, deployed across a vast coverage area spanning thousands to millions of square kilometres. Cellular communication relies on a network of base stations, each serving users in a part of the coverage area. Each tower's coverage area is referred to as a cell, hence the name cellular communication (Fig. 1).

1.1 Background: 1G–5G

Adoption of cellular communication started in the 80s with Advanced Mobile Phone System (AMPS), an analog system where each user's voice was frequency modulated and transmitted over a user-specific 30 kHz bandwidth channel allocated for each phone call. AMPS and other analog wireless cellular communication systems of that era are retronymously referred to as the first generation of cellular—1G.

This was followed in the 90s by the second generation (2G) technologies such as GSM and IS-95. In a time-domain multiple access (TDMA) system like GSM[13], each user was allocated periodic slices of time with a 200 kHz bandwidth, during which digitized and compressed voice was transmitted using digital communication techniques such as GMSK modulation and convolutional codes. Signal processing techniques such as equalizers were typically used to combat channel impairments such as multipath intersymbol interference. In IS-95, a spread spectrum code division multiple access (CDMA) system[11], each user was allocated a unique code. Digitized and compressed voice was convolutionally encoded and BPSK or QPSK modulated before being 'spread' across a much wider bandwidth (1.25 MHz) using the user-specific code. A Rake receiver[11], comprised of multiple 'fingers', was used to recover the signal in the presence of multipath. Each finger demodulated the transmit symbol received with a specific multipath delay. The outputs of all the fingers were then combined, thereby harnessing the total energy received through the multiple paths between the transmitter and the receiver. 2G systems initially supported voice and short-message services. Through evolution to GPRS and EDGE, they added support for data rates on the order of tens up to low hundreds of Kbps.

3G arrived in the early 2000s. Standardized by the Third Generation Partnership Project (3GPP), Wideband CDMA (WCDMA)[9], as its

1G: Analog Cellular Communication Systems are retronymously referred to as 1G.

[1] *Sunnyvale, CA, USA.* **psubrahmanya@gmail.com*

Figure 1: Cellular communication network.

OFDMA: The structure of OFDM allowed user transmissions to be separated in the frequency domain in addition to the time domain. This technique is referred to as OFDMA.

Shared channel: A physical layer shared channel allowed the channel to be sliced in the time and code domains.

name suggests, used a 5 MHz bandwidth, over three times wider than that of IS-95. In addition to voice, WCDMA and its evolution HSDPA supported high-speed data at rates up to 14.4 Mbps. Contemporaneously, through a parallel standard organization 3GPP2, IS-95 evolved into cdma2000 1x, primarily for voice, and EV-DO[10] for high-speed packet data transport[12]. 3G introduced a number of new physical layer techniques to Cellular communications such as Turbo codes[14,15], Hybrid ARQ[21], and Higher Order QAM Modulation. A physical layer **shared channel** allowed the channel to be sliced in the time and code domains. Scheduling techniques such as proportional-fair scheduling allowed multiple users to share access to the channel efficiently while maintaining fairness. High-speed data transport also necessitated operation at higher SNRs. Equalization was required to mitigate inter-chip interference caused by multipath. 3G evolution continued to achieve higher data rates through techniques such as Multi-Carrier and MIMO[20].

4G was introduced late in the 2000's. Also known as LTE (Long-Term Evolution)[8], the new standard was based on Cyclic Prefix-Orthogonal Frequency Division Multiplexing (CP-OFDM)[19] on the downlink and single-carrier FDMA[18] on the uplink. LTE supported bandwidths ranging from 1.4 MHz to 20 MHz. Wider bandwidths allowed for an increase in achievable data rates. Using a cyclic prefix between adjacent symbols, CP-OFDM offers an elegant solution to the problem of inter-symbol interference caused by multipath. At the transmitter, a set of QPSK or QAM symbols to be transmitted over one OFDM symbol is used to modulate the gain and phase of a set of tones chosen to be orthogonal over the symbol duration. This operation can be performed efficiently using a Fast Fourier Transform (FFT). The resulting modulated tones are combined and a trailing portion of the resulting time domain signal is added as a prefix to the signal. As long as the multipath delay spread is less than the duration of the cyclic prefix, there is a window of time during which transmissions from adjacent symbols do not overlap, thus avoiding ISI. By transforming the signal using an FFT, equalization can be performed in the frequency domain with far less complexity than a time-domain equalizer, whose taps would increase proportionately with bandwidth. The physical layer shared channel concept from 3G carried over. The structure of OFDM allowed user transmissions to be separated in the frequency domain in addition to the time domain. This technique is referred to as **OFDMA**.Over the course of a few years, LTE data rates steadily increased into the Gigabits per second range through techniques such as 256-QAM modulation, four-layer MIMO, and aggregation of multiple LTE carriers. Specialized features were added to support use cases such as the Internet of things (NB-IOT) and vehicular communication (C-V2X). While LTE was initially designed to operate in licensed bands, later releases of the standard also added the ability to deploy in unlicensed bands.

The progression of data rates from 1G through 5G is shown in Fig. 2.

Following 1G through 4G, 5G NR (New Radio) has been developed over the last few years to continue the evolution of Cellular Communication technologies. In contrast to prior generations which focused primarily on use cases dominated by the (smart)phone, the design of 5G NR from the very beginning has considered a variety of use cases such as smartphones, fixed wireless access (FWA), Industrial IOT, and other vertical application domains. Also, the design of 5G has considered deployment in a much richer set of frequency bands, going well beyond 1G through 4G which were primarily deployed in bands below 3 GHz. In the subsequent sections, we will describe how these considerations have guided the design of 5G.

1.2 Outline of Rest of the Paper

In Sect. 2, we outline the overall structure of the 5G physical layer. In Sect. 2.1, we describe the overall structure of the 5G NR physical layer, followed by a description of the time and frequency structure of the 5G waveform in Sect. 2.1.3, and

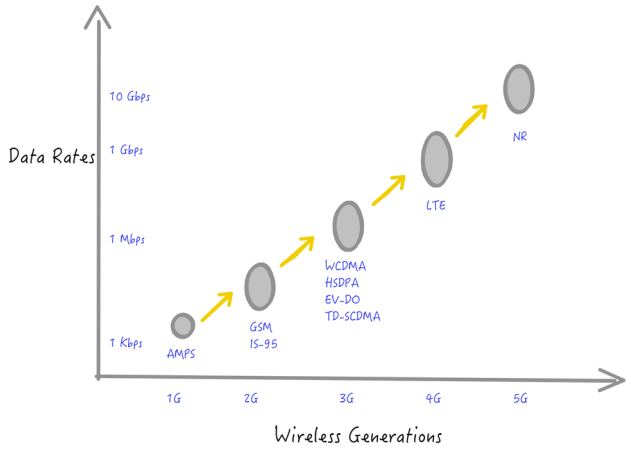

Figure 2: Cellular communication network.

a description of individual physical channels in Sect. 2.1.4.

In Sect. 3, we elaborate on some of the guiding principles of the 5G physical layer: operation in high-frequency bands (Sect. 3.1), lower latency (Sect, 3.3), flexibility (Sect. 3.4), future extensibility (Sect. 3.5), and coexistence with LTE (Sect. 3.6).

Finally, we describe future directions beyond the first 5G NR Release in Sect. 4 and conclude the paper in Sect. 5.

2 The 5G NR Physical Layer

In this section, we review the overall structure and some of the principles guiding the physical layer of 5G NR[1–6]. The treatment is pedagogical in nature. Our focus is on communicating the underlying concepts, not on an exhaustive treatment of the details. The 3GPP specification supports a multitude of formats, configurations, and deployment possibilities. In the interest of readability, we cover only a small subset of this vast space of possibilities, referring the reader to the specifications themselves or other references for a complete description.

2.1 Overall 5G NR Physical Layer Structure

5G NR operates within the same overall over-the-air procedural paradigm followed by the preceding generations. A number of over-the-air physical channels are defined, each occupying a set of time and frequency resources, and designated for a specific purpose.

2.1.1 Overall Over-the-Air Paradigm

Any cellular system consists of a network of base stations (referred to as gNodeB's in 5G NR terminology) and mobile devices [referred to as User Equipment (UE)]. In any cellular system, a UE discovers the availability of a suitable network using synchronization signals transmitted by the network. It then reads system information broadcast by the network to determine how to access the system. Typically, a pool of uplink resources in frequency and time is reserved for access and the UE follows a randomized procedure to select resources within this pool for the initial transmission, along with a backoff mechanism for contention resolution. After the initial communication is established, data transmission to (from) the UE takes place over a downlink (uplink) shared channel. A downlink control channel signals the presence and format of downlink data transmissions to the UE and indicates the time and frequency resources and format that the UE should use for its uplink transmissions. The UE uses an uplink control channel to indicate the receipt of downlink transmissions, to provide feedback on the state of the downlink channel, and to request resources for uplink transmissions. **The wireless channel** between the transmitter and receiver transforms the signal through a number of mechanisms such as delay spread due to multi-path, time/frequency selective channel fading, Doppler spread, doppler shift, shadowing, blockage, etc. Additionally, typical wireless transmitters and receivers exhibit a number of non-idealities such as frequency and timing drift due to clock skew, phase noise, etc. Reference signals are transmitted, so that the receiver can estimate the characteristics of the channel and receiver non-idealities

The wireless channel: The wireless channel between the transmitter and receiver transforms the signal through a number of mechanisms.

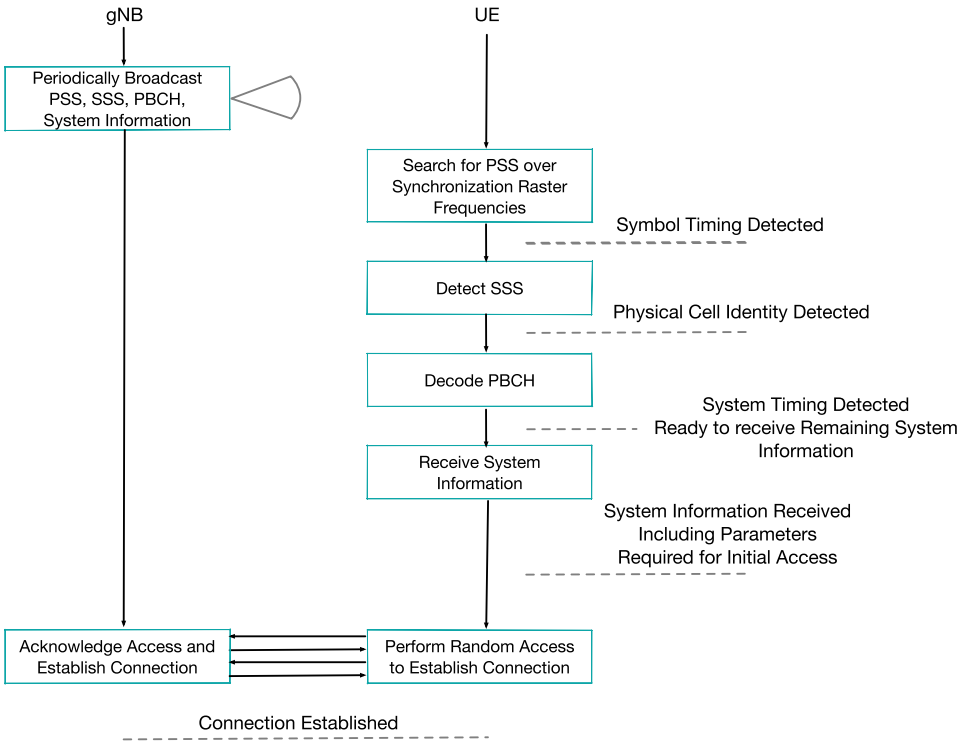

Figure 3: Initial access.

and account for them in receiving the signal. Some channel characteristics may also be fed back to the network for use in optimizing transmissions.

2.1.2 Mapping Overall Paradigm to 5G NR

5G NR includes physical channels providing all of the above functionalities. A Primary Synchronization Signal (PSS) is provided to enable system discovery, along with a Secondary Synchronization Signal (SSS) to disambiguate multiple cells that transmit the same PSS sequence. A subset of system information is sent over the Physical Broadcast Channel (PBCH), which also includes a pointer to the resources carrying the remaining minimum system information required to access the network. PSS, SSS, and PBCH, are transmitted in a time-contiguous sequence, and are referred to as a Synchronization Signal Block (SSB) A Random Access Channel (RACH) enables random access to the network. Downlink and Uplink data transmission take place over a Physical Downlink Shared Channel (PDSCH) and Physical Uplink Shared Channel (PUSCH) respectively, with accompanying control information on a Physical Downlink Control Channel (PDCCH) and Physical Uplink Control Channel (PUCCH) correspondingly. A number of reference signals are provided: Demodulation Reference Signals (DM-RS) for channel estimation, Tracking Reference Signal (TRS) for estimating and tracking various channel characteristics, Phase Tracking Reference Signal (PTRS) for estimating phase noise, Channel State Information-Reference Signal (CSI-RS) to enable downlink channel state estimation for feedback to the network, and a Sounding Reference Signal (SRS) for a similar purpose on the uplink. The above channels are further elaborated in the following sections.

Random access: The UE avails of the synchronization channels and broadcast system information to acquire and perform a random access procedure to establish a connection with the network.

In keeping with the overall over-the-air paradigm of cellular communication, as shown in Fig. 3, the UE avails of the synchronization channels and broadcast system information to acquire and perform a **random access** procedure to establish a connection with the network. The steps of the random access procedure are elaborated in Fig. 4.

2.1.3 Time and Frequency Structure

Figures 5, 6 depict an example realization of the overall time and frequency structure of the 5G NR waveform. 10 ms frames are divided into 1 ms subframes, which in turn are sub-divided

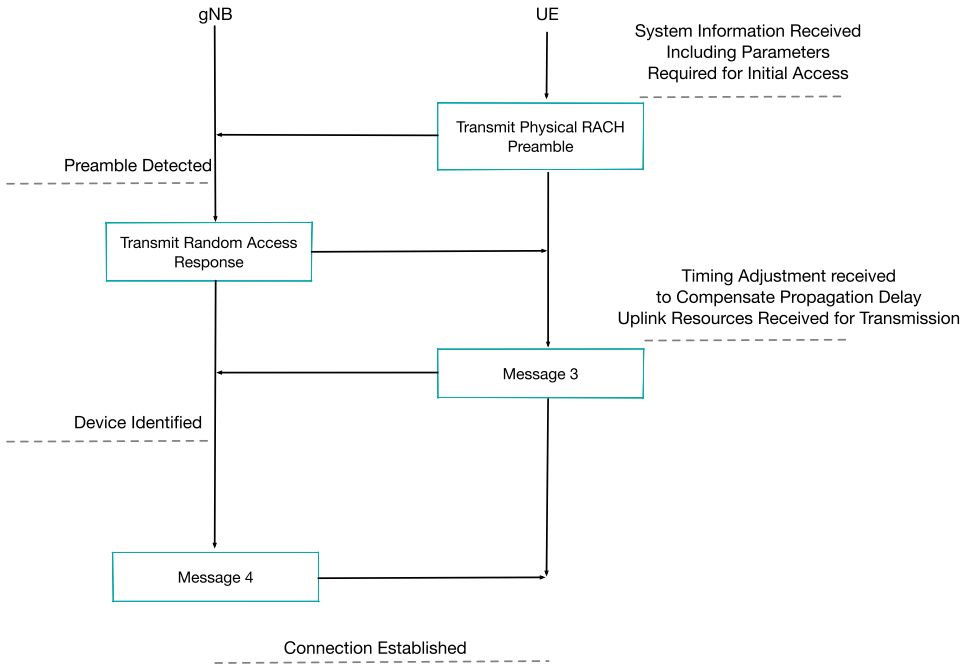

Figure 4: Random access procedure.

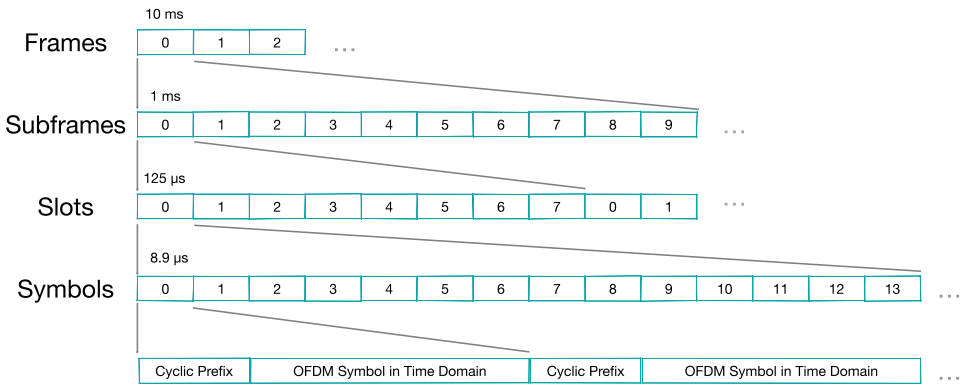

Figure 5: Time-domain structure.

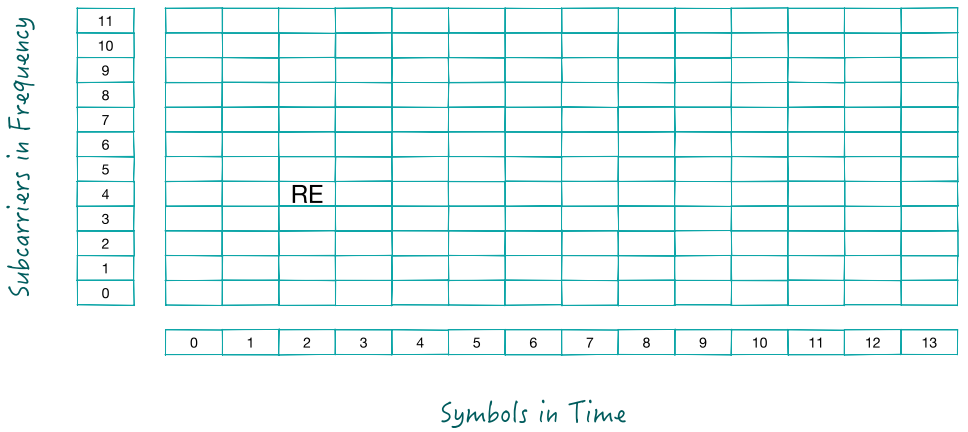

Figure 6: Frequency-domain structure of a resource block.

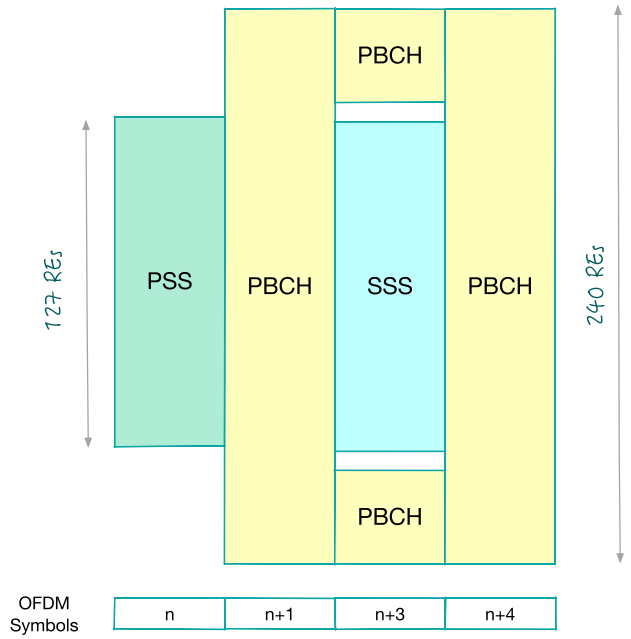

Figure 7: Structure of SSB.

LDPC: Some of the changes introduced in NR include the use of LDPC codes for the PDSCH and Polar codes for PDCCH.

into one or more slots. The number of slots within a subframe depends on the spacing between subcarriers. The example shown corresponds to a subcarrier spacing of 120 kHz. A slot is made up of 14 symbols. Each symbol includes a cyclic prefix to enable OFDM processing at the receiver.

In the frequency domain, information is modulated over resource elements (RE), each of which corresponds to a single OFDM tone or subcarrier transmitted over a single-symbol duration. Resource Blocks (RB) are defined comprising 12 REs in frequency and 14 symbols in time. In the example shown, with a subcarrier spacing of 120 kHz, a single RB occupies 1.44 MHz in frequency. A 5G NR signal is composed of a number of RBs, such that the total bandwidth occupied by all the RBs fits within the available system bandwidth.

Within this overall time and frequency structure, each of the physical channels defined above is mapped to a specific set of REs.

2.1.4 Structure of Individual Physical Channels

Figure 7 shows the structure of an SSB. An SSB occupies 4 symbols in time and 240 subcarriers in frequency. There are 3 possible PSS sequences and 336 SSS sequences, thus allowing 1008 cell identities to be distinguished.

Figure 8 depicts the processing steps involved in transmitting a PDSCH, while Fig. 9 depicts the steps for a PDCCH. Some of the changes introduced in NR include the use of **LDPC** codes[16] for the PDSCH and Polar codes[17] for PDCCH. The transition from Turbo codes in LTE to LDPC codes in NR was primarily driven by the high-throughput demands of 5G NR. The structure of LPDC codes allows for decoder implementations with more parallelism, making them attractive for very high-throughput scenarios. It has been shown that LDPC codes can achieve a better performance with fewer computations than Turbo codes. The LDPC code structure in NR was defined to support effective incremental redundancy hybrid ARQ, along with a wide range of block lengths and coding rates.

Figure 10 shows an example of resource mapping of corresponding PDCCH and PDSCH. The PDCCH can occupy up to three symbols in time. PDCCH and PDSCH have corresponding DM-RSs to allow for the respective channel estimation.

CSI-RS and SRS are reference signals used for sounding the channel, measuring interference, as well as supporting beam management and mobility in NR. CSI-RS is used on downlink transmissions from the gNodeB to the UE and SRS is used on uplink transmissions from the UE to the gNodeB. Tracking reference signal (TRS) can be viewed as a collection of single-port CSI-RS with a particular

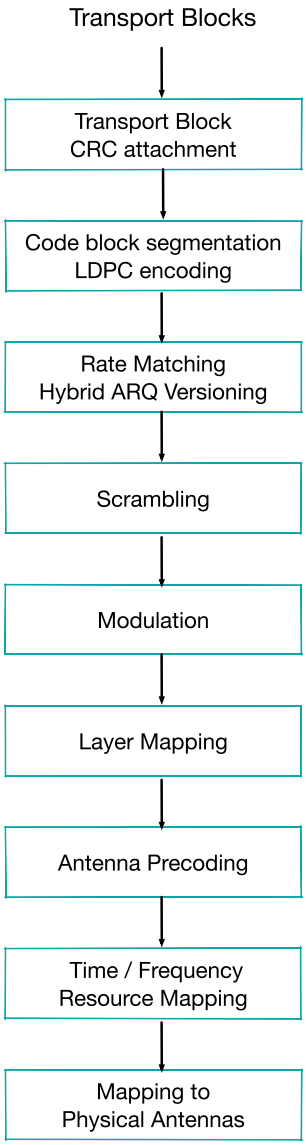

Figure 8: PDSCH processing steps.

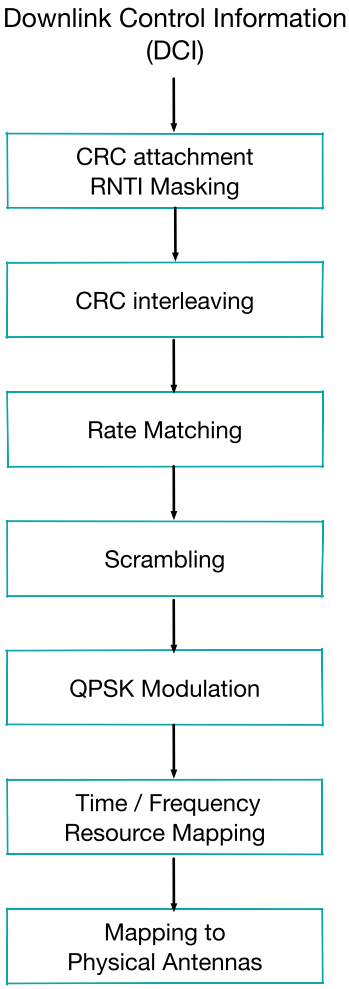

Figure 9: PDCCH processing steps.

structure which is used for estimating various channel characteristics as well as correcting for timing and frequency errors caused by imperfect oscillators.

Phase tracking reference signal (**PTRS**) is a newly introduced reference signal in 5G NR for mitigation of the impact of phase noise from the transmit and receive oscillators. The impact of phase noise is more pronounced in higher frequency bands. 5G NR supports PTRS transmission in both downlink and uplink, and can be applied for both CP-OFDM and SC-FDM operation. Phase noise will impact OFDM and SC-FDM systems by rotating symbols on different subcarriers with a common phase error (CPE), as well as causing inter-carrier interference. CPE can be estimated and compensated using PTRS. PTRS is transmitted only within the scheduled resource blocks for PDSCH and PUSCH transmission. Given the properties of phase noise, PTRS is sparse in the frequency domain and denser in the time domain. The reference signal structure is configurable in terms of density in frequency and time domain. This flexibility allows the system to configure PTRS depending on the carrier frequency, subcarrier spacing, and quality of oscillators. The temporal density can be dynamically adjusted based on the modulation and coding rate used in transmission to balance the overhead caused by transmission of PTRS with the impact of CPE residual error. Figure 11 illustrates two different configurations of PTRS.

PTRS: Phase tracking reference signal (PTRS) is a newly introduced reference signal in 5G NR for mitigation of the impact of phase noise from the transmit and receive oscillators.

Figure 12 shows the processing steps involved in transmitting the PUSCH. Some changes in the NR uplink include the introduction of CP-OFDM support in addition to DFT-spread OFDM and

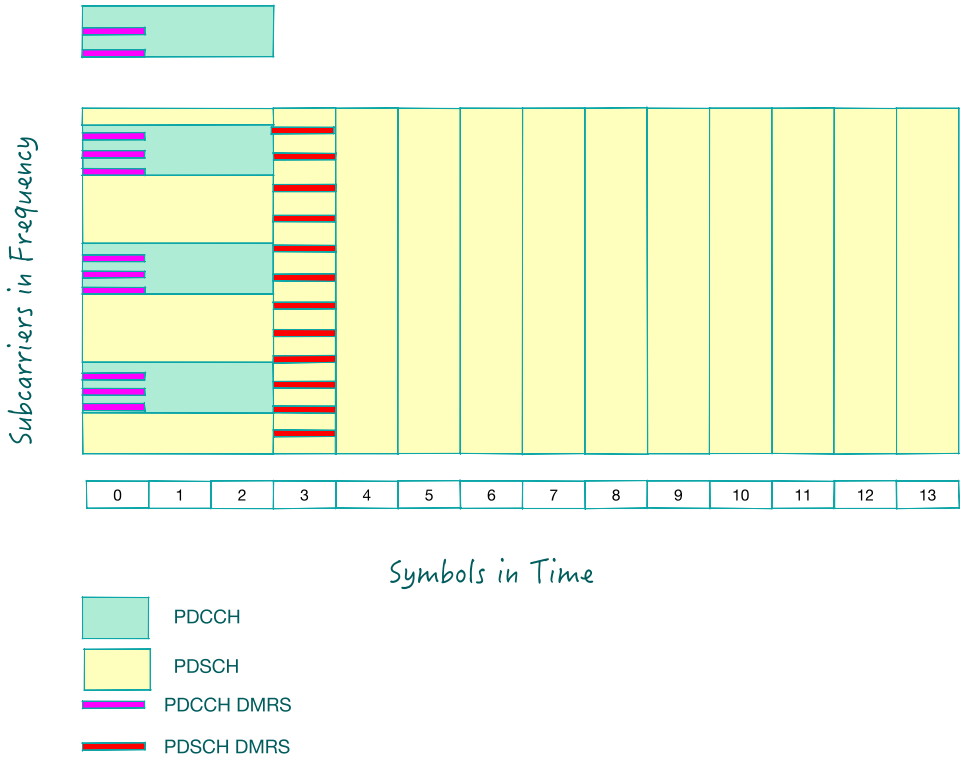

Figure 10: Example of resource mapping of PDCCH and PDSCH.

support for $\frac{\pi}{2}$-BPSK as an additional modulation scheme.

3 Guiding Principles of the 5G Physical Layer

In this section, we elaborate on the guiding principles behind the 5G NR Physical layer.

3.1 Operation in High-Frequency Bands

Higher frequency: Higher frequency bands offer a large amount of usable spectrum, enabling higher data rates and lower latencies.

Beams: Without additional receiver/transmitter techniques, the coverage of a millimeter wave cell would shrink to an unacceptably small cell radius, necessitating an expensive deployment of a large number of additional cell sites to achieve acceptable coverage. To avoid this and to mitigate propagation loss, both the network and the UE employ the use of 'beams'

The development of 5G has considered the need for deployment in a variety of frequency bands (Fig. 13). As opposed to 1G through 4G which operated in bands below 3 GHz, 5G has been designed to allow for deployment in frequency bands as high as 52.5 GHz in Release 15 with even higher frequency bands under consideration in future releases. **Higher frequency** bands offer a large amount of usable spectrum, enabling higher data rates and lower latencies.

Frequency bands below 6 GHz are colloquially referred to as "sub-6" GHz bands, while the higher frequency bands in the range of 24 GHz and 52.5 GHz are referred to as "millimeter wave" frequency bands, indicating the wavelengths of signals in these bands[7]. The 3GPP specifications refer to frequency bands below 6 GHz as FR1 (Frequency Range 1) and frequency bands between 24GHz and 52 GHz as FR2 (Frequency Range 2).

Operation in millimeter wave frequency bands necessitates the development of a number of special techniques. Radio signal transmissions in millimeter wave frequency bands experience higher propagation losses. For instance, based on the $\frac{1}{f^2}$ loss trend, a signal at 39 GHz would experience 16 dB greater propagation loss than a signal at 6 GHz. Without additional receiver/transmitter techniques, the coverage of a millimeter wave cell would shrink to an unacceptably small cell radius, necessitating an expensive deployment of a large number of additional cell sites to achieve acceptable coverage. To avoid this and to mitigate propagation loss, both the network and the UE employ the use of '**beams**' (Fig. 14).

As depicted in Fig. 14, the use of beams to mitigate propagation loss is analogous to the focusing of a beam of light from a flashlight to enable it to propagate further. Beams are formed using antenna arrays at the transmitter and receiver: transmit beams, where the signal directed to each of the antennas in the array is modified in phase so as to add constructively in the desired direction

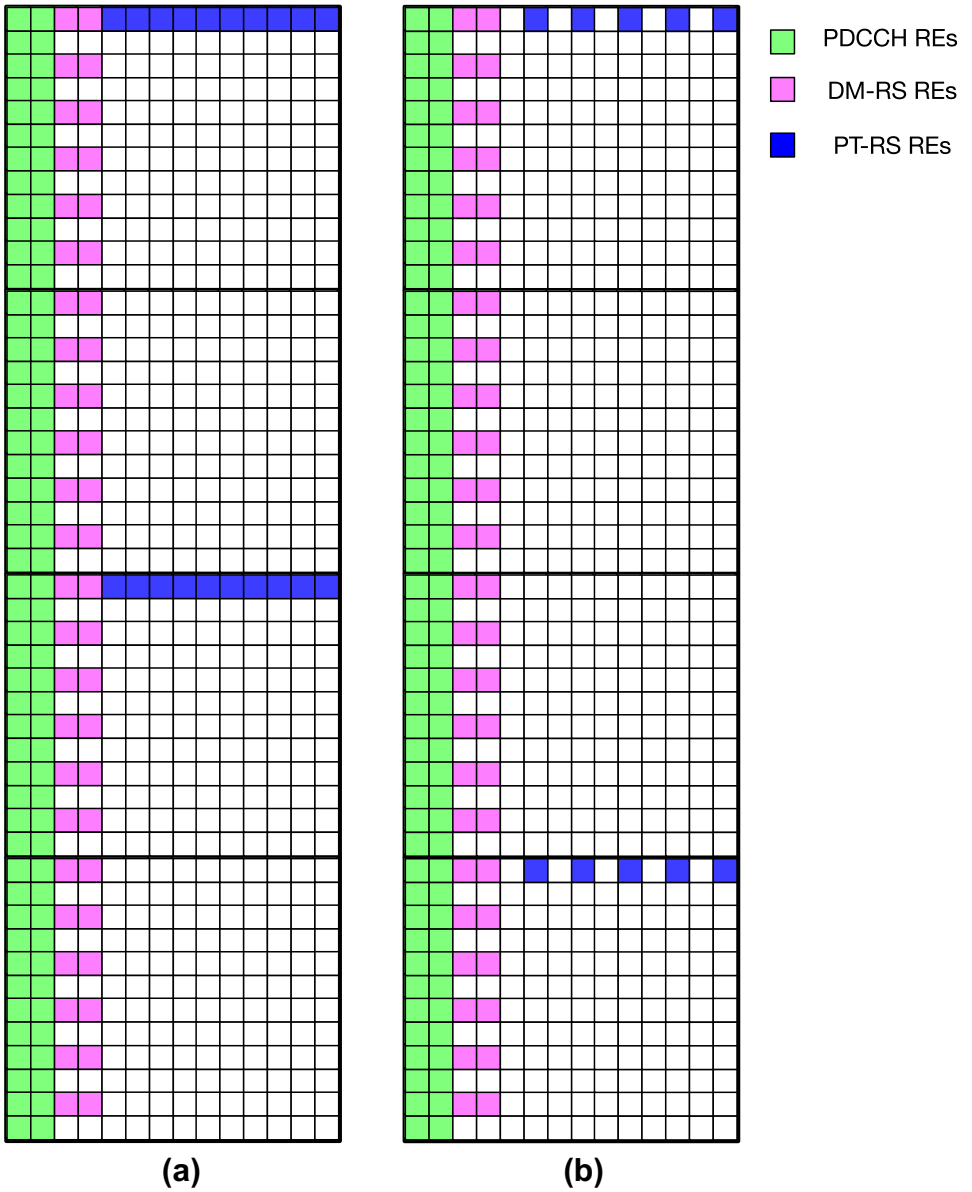

Figure 11: **a** PTRS configuration for large MCS and medium RB allocation PDSCH. **b** PTRS configuration for lower MCS and large RB allocation PDSCH.

of the receiver and receive beams, where the signal received at each of the antennas is modified in phase before being combined, such that the signal component received from the desired direction of the transmitter is selectively amplified. This is depicted in Fig. 15 for the case of receive beamforming. In the figure, the signal to be received arrives from a direction making an angle θ with the array. The signal reaches each antenna after traversing an additional distance $d\cos\theta$ relative to its neighbour to the right. For a tone at a carrier frequency f, the additional delay incurred in traversing this distance leads to an additional phase shift $\phi = 2\pi f \frac{d\cos\theta}{c}$, where c is the speed of light. If the received signal from each antenna is provided an additional phase shift of ϕ relative to its neighbour to the left, the signals from all the antennas will combine constructively.

A waveform structure and beam identification procedures are defined, such that both the gNodeB and UE can identify the appropriate beams to use for transmission and reception on the downlink and uplink.

To enable the initial access to the system, the gNodeB transmits synchronization signals (SSBs) using each of its beams, one after the

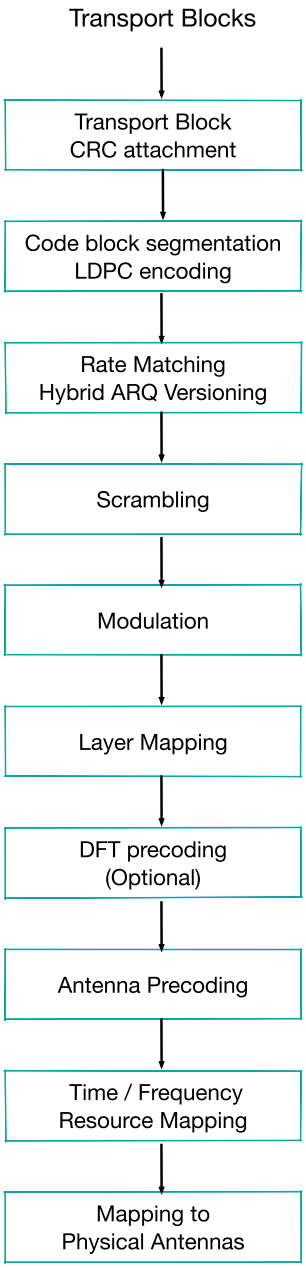

Figure 12: PUSCH processing steps.

Beam correspondence: On the uplink, the UE may determine its transmit beam based on the optimal receiver beam it selected for the downlink, relying on channel reciprocity. This process, in concert with a similar determination at the network, is referred to as beam correspondence.

other (Fig. 16). The UE makes measurements using multiple receive beams and identifies the synchronization signal that it can receive with the highest quality (in terms of signal strength or SNR). Each synchronization signal is associated with corresponding time and frequency resources and preamble sequences with which the UE can initiate access. By initiating random access during a slot associated with the highest received quality synchronization signal, the UE implicitly communicates to the network the gNodeB beam that is most suitable for communication with it.

After suitable beams are initially chosen, they need to be continually updated to keep up with UE motion. SSBs can be periodically monitored. In addition, the network may transmit CSI-RS signals to enable beam measurements. On an ongoing basis, the UE performs measurements using different SSB and CSI-RS signals using different UE receive beams. Using these measurements, the UE can adapt its own choice of receive beams, while beam measurement reports sent to the network enable it to select the optimal beams for transmission to the UE.

On the uplink, the UE may determine its transmit beam based on the optimal receiver beam it selected for the downlink, relying on channel reciprocity. This process, in concert with a similar determination at the network, is referred to as **beam correspondence**. Alternately (or additionally), the network may configure the UE to transmit SRS using multiple transmit beams. SRS measurements can then be used to determine the optimal UE transmit beam.

Enabling operation in high frequency bands is critical to utilizing a much larger amount of spectrum for cellular communications.

3.2 Wider Bandwidths

Achieving higher throughput by utilizing the large amounts of spectrum available, especially in the higher frequency ranges, requires techniques to receive and transmit wide bandwidth signals. In LTE, the largest system bandwidth supported was 20 MHz. Utilizing larger amounts of spectrum required carrier aggregation, even for contiguous spectrum. Each component carrier carries a certain amount of overhead. Furthermore, procedures to add and remove component carriers introduce delay and slow down responses to system conditions.

As shown in Fig. 17, 5G NR allows system bandwidths as wide as 100 MHz in FR1 and 400 MHz in FR2. Additionally, as in LTE, carrier aggregation can be used to aggregate even larger amounts of spectrum. The widest bandwidths are supported with higher subcarrier spacings and smaller slot durations. This avoids scaling implementation complexity linearly with bandwidth. For instance, the same FFT size can support twice the bandwidth with a subcarrier spacing twice as large. Similarly, some of the processing buffers scale with bandwidth–duration product; halving the slot duration while doubling the bandwidth avoids having to increase the buffer size.

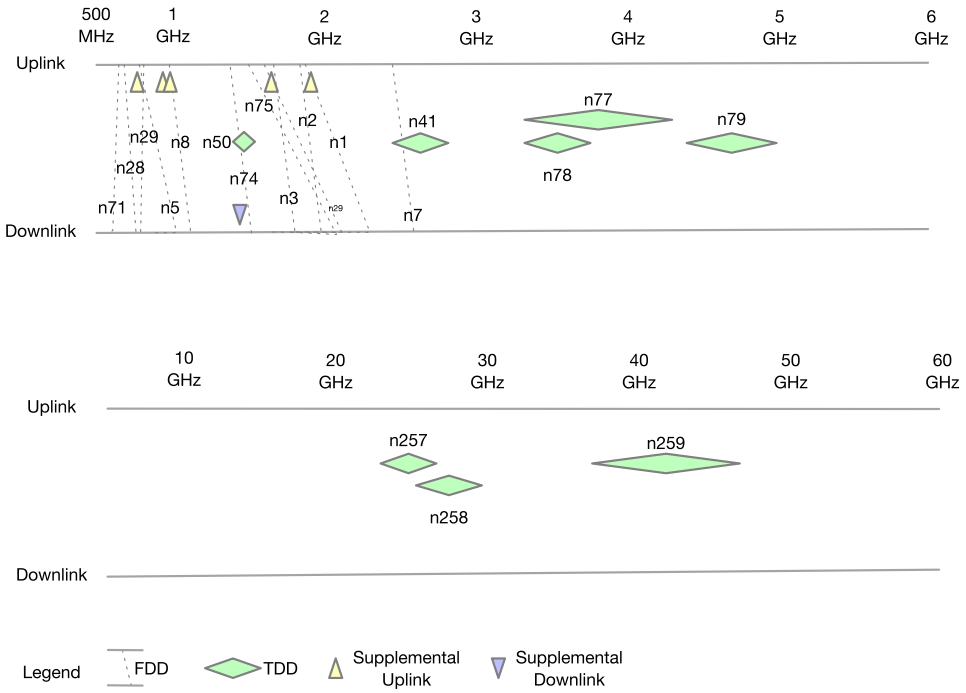

Figure 13: 5G frequency bands.

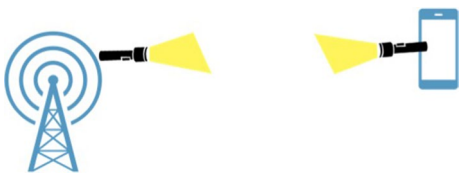

Figure 14: Beamforming concept—flashlight analogy.

5G NR also enables lower complexity UEs, such as those used in low-power IOT devices, to limit their support for wide bandwidths. This is accomplished through the introduction of bandwidth parts. A bandwidth part constitutes a portion of the full system bandwidth. The network assigns smaller bandwidth parts to lower complexity UEs, allowing them to coexist with higher

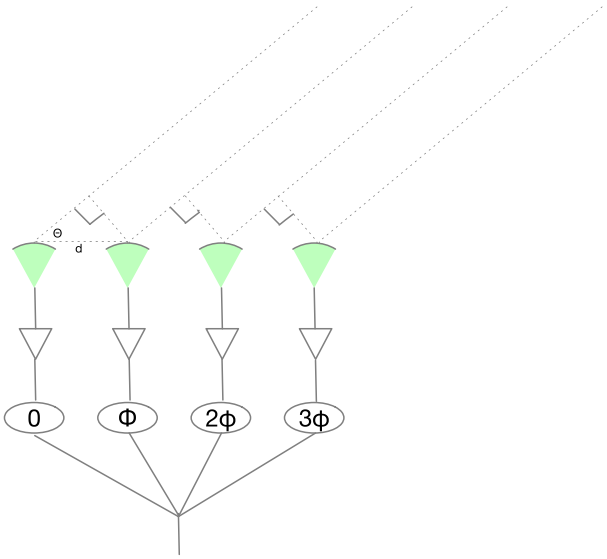

Figure 15: Receive beamforming.

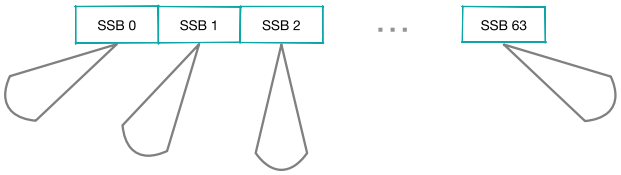

Figure 16: Synchronization signals and beams.

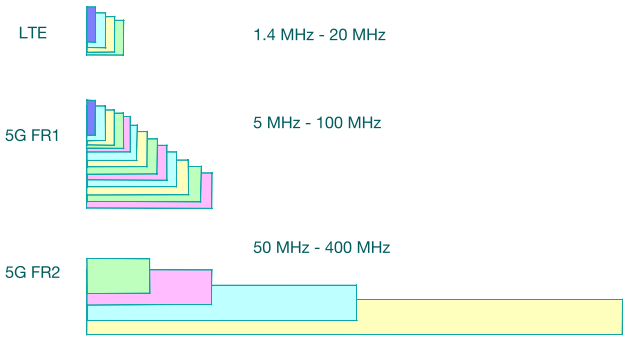

Figure 17: Bandwidths.

complexity UEs within the same wide system bandwidth. Smaller bandwidth parts can also be assigned even to full-complexity UEs. In system states such as Idle, or during periods of low data activity, the UE does not benefit from having access to the full system bandwidth. In such cases, the network can restrict the UE to a smaller bandwidth part. A narrower bandwidth consumes less power, since receiver and transmitter processing typically scale monotonically with bandwidth. The lower processing requirements can be met in lower power states within the UE, thereby extending battery life.

3.3 Lower Latency

At the application level, support of higher data rates translates into a shorter time to download or upload large amounts of data. To get the full benefit of these higher data rates, it is important to simultaneously reduce latency even for smaller transmissions. Otherwise, the effective reduction in download/upload time may be lower-limited by the amount of time required to complete various other communication steps such as establishing client–server connections, protocol synchronization, etc. Limited buffer sizes in various entities may also mean that the increase in bandwidth–delay product prevents taking full advantage of the increase in bandwidth.

Furthermore, there are applications that require lower latency, but do not require large amounts of data. Examples of such applications include control applications where it is important to send system measurements and control messages very quickly, so that reaction time can be minimized.

5G NR achieves lower latencies through a number of mechanisms. The smaller slot sizes (as low as 125 ms) are one aspect of this. Additionally, there are multiple symbol positions within a slot where transmissions can start or stop. This reduces latency even further, since incoming data do not need to wait for the start of the next slot before being scheduled for transmission, and smaller transmissions can be completed within a fraction of slot.

The signal structure within a slot further reduces latency. In LTE, the structure of the PDCCH in a given subframe is unknown to the UE prior to that subframe. The network can configure the PDCCH to occupy anywhere from 1 to 3 symbols. The UE needs to first process a Physical Control Format Indicator Channel (PCFICH) to determine the structure of the PDCCH. It then decodes the PDCCH before decoding the PDSCH. 5G NR reduces delay in PDCCH processing by eliminating the PCFICH. The structure of the PDCCH is known in advance to the UE, so it can start processing the signal as soon as it is received. PDSCH reception in LTE was also slowed down by the need to receive

Table 1: Examples of NR configurability.

Parameter	FR1 choices	FR2 choices
System Bandwidth	{1...10} × 10 MHz,	{1...4} × 50 MHz
	5, 15, 25 MHz	
Subcarrier spacing	{1, 2, 4} × 15 kHz	{4, 8, 16} × 15 kHz
Slot duration	1 , 0.5 , 0.25 ms	0.25 , 0.125 , 0.063 ms
Cyclic prefix duration	4.7 μs, 2.3 μs, 1.2 μs	1.2 μs, 0.59 μs, 0.29 μs
Slot formats	Several (see examples)	Several (see examples)
ACK/NAK Turnaround Time	Several	Several
DM-RS positions	{slot, data}-aligned	same
	×{ 1 symbol, 2 symbol }	
	×{ Type 1, Type 2 }	
	×{ 1, 2, 3, 4 } DM-RS	

demodulation reference signals (CRS or UE-RS). 5G NR includes modes where DM-RS signals can be placed at the start of the PDSCH. This aligns signals with their processing order in the UE, thereby reducing processing delay.

The latency between the receipt of control information on the downlink and transmission on the uplink has been reduced. Similarly, the latency between the receipt of the PDSCH and transmission of acknowledgments on the uplink has been reduced.

5G NR also allows for urgent transmissions starting in the middle of a slot to pre-empt transmissions that may have commenced earlier in the slot. Pre-emption indication signals are included, so that the UE whose reception was disrupted by a new transmission can account for the corrupted symbols when processing. Receiver acknowledgments sent at the granularity of code block groups (portions of a transport block) also allow the pre-empted transmissions to be retransmitted and recovered efficiently.

5G NR also includes a better pipelining of processing between the RLC, MAC, and physical layers to reduce latency. Additionally, resources can be pre-configured on the uplink, so that there is no need to incur delay in requesting resources first before transmitting data.

3.4 Flexibility

As mentioned earlier, 5G NR supports a multitude of configurations and formats. This allows each deployment to be customized for its needs.

Table 1 shows some examples of high-level system-level configurability in NR. System bandwidths can be tuned to the available spectrum. Larger subcarrier spacing can be selected for operation at higher frequency bands where phase noise is larger to reduce the impact of phase noise on inter-carrier interference, while smaller subcarrier spacing such as 15 kHz is more appropriate for lower frequency bands where coexistence with LTE is important. Cyclic prefix duration can be tuned to the delay spreads anticipated in the specific deployment. Note that not all combinations of parameters are valid. For instance, a larger subcarrier spacing implies a smaller symbol duration, and hence a shorter cyclic prefix. NR allows for support of 240 kHz SCS for SSB in a system where other channels use a subcarrier spacing of 120 kHz.

5G NR supports both FDD and TDD using a common set of frame structures. In TDD, each slot can be flexibly configured into uplink and downlink symbols. This allows the network scheduler to tune the uplink-to-downlink ratio to match live operating conditions. A wide variety of slot formats are supported as seen from the examples in Fig. 18. Symbols within a slot can be configured to be Uplink-only, Downlink-only, or Flexible. Flexible symbols can be used for Downlink or Uplink or used as a guard period. Downlink control information sent by the network determines the usage of Flexible symbols by the UE on a slot-by-slot basis. It must be noted that in TDD systems, a variety of interference mechanisms exist between transmissions from different gNodeBs and/or different UEs. Deployment

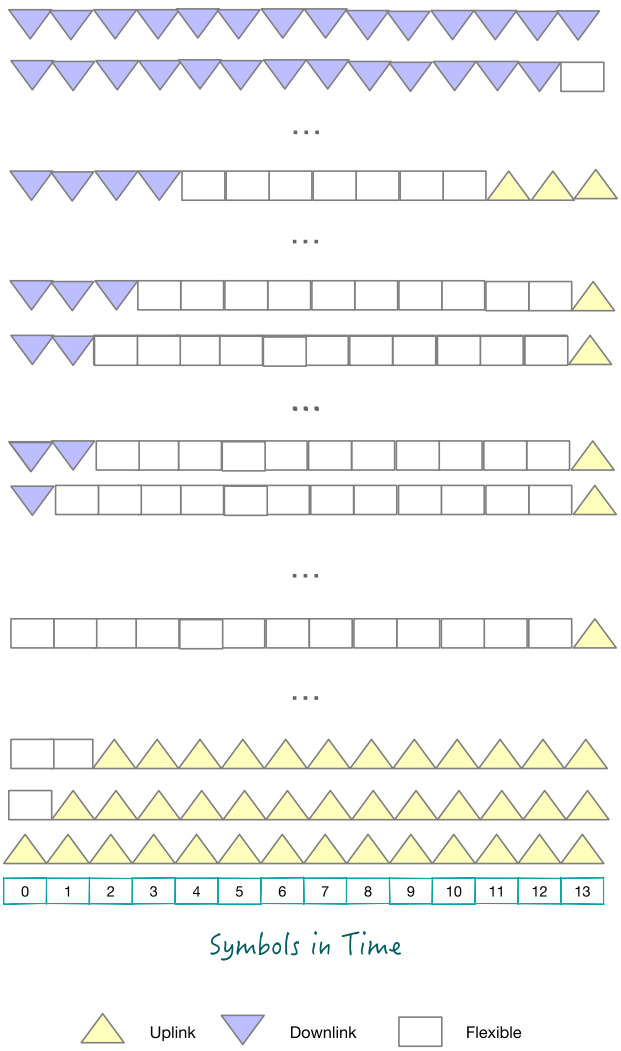

Figure 18: Slot format examples.

configuration needs to be considered in using the available flexibility. For instance, it would not be wise for a gNodeB to alter slot formats on a slot-by-slot basis without coordination with nearby gNodeBs.

The RS structure in 5G NR is quite flexible. In addition to supporting DM-RS positions at fixed offsets from the start of a slot, the DM-RS can also be sent at the start of data transmission. This is useful for data transmissions that occupy only a small portion of the slot. The number of DM-RS occasions within a slot is also configurable. Multiple DM-RS symbols within a slot are useful to allow tracking of a rapidly changing channel in high mobility environments, while a single DM-RS symbol is preferable in low mobility environments, freeing other symbols for data transmission. The density of DM-RS signals in the frequency domain is also configurable—type 1 DM-RS has a higher density of DM-RS signals, which is advantageous when the channel frequency selectivity is high, while type 2 DM-RS has a lower density, allowing the multiplexing of a larger number of DM-RS signals, advantageous when multi-user MIMO transmissions are scheduled to a large number of devices using the same time-frequency resources.

Coresets: The control region is defined in terms of time and frequency limited resource regions called Coresets.

Another example of flexibility in 5G NR is the structure of the control channel. The control region is defined in terms of time and frequency limited resource regions called **Coresets**. Coresets can be tuned to the capabilities of different UEs. For instance, UEs capable of processing a smaller bandwidth can be assigned Coresets with narrower frequency extent. Unused control resources within a Coreset can also be reused for PDSCH transmission. This allows for improved system efficiency.

Yet, another example of flexibility is in the turnaround time between the receipt of the PDSCH and the transmission of an acknowledgment on the uplink. In LTE, this time was fixed at 3 ms nominally (timing adjustments could further reduce this number by up to 0.66 ms). In 5G NR, the turnaround time is flexible and is indicated by the DCI. The flexibility in scheduling the uplink acknowledgment allows the scheduler to take advantage of the several slot formats available in NR to schedule the transmission of acknowledgments in a manner that is consistent with the goal of optimizing system-wide operation.

3.5 Future Extensibility

5G NR has been designed such that it can be extended in the future with further improvements. The same spectrum can be shared in the future between 5G NR users and users of future versions or generations.

To that end, the signals that 'must' be transmitted in a 5G NR system have been kept to the bare minimum. This leaves most time and frequency resources to be repurposed in future releases or generations. Some examples of this principle in action follow.

In LTE, a cell-specific reference signal was transmitted even in the absence of data transmissions, to allow new incoming users to perform channel estimation to receive overhead signals. In NR, each physical channel includes its own demodulation reference signals. Thus, without a transmission, the corresponding reference signals would not be transmitted either.

In LTE, the PSS and SSS signals must occur with a 5 ms periodicity. In NR, the SSB periodicity can range from 5 ms all the way to 160 ms.

Another way to support future extensibility is to allow the time or frequency resources used for any physical channel to be configurable. Some examples of this principle in action follow.

In LTE, all subcarriers in the first 1–3 symbols of the subframe were reserved for PCFICH/PDCCH. In contrast, in NR, the control region is defined in terms of Coresets, and can flexibly occupy portions of the time and frequency resources available.

In LTE, the uplink follows a synchronous hybrid ARQ protocol where the retransmission needs to occur at a specific point in time. In NR, there is more flexibility in rescheduling the retransmission.

3.6 Coexistence with LTE

One of the design principles of 5G NR has been the ability to **coexist** within the LTE physical layer structure to allow for sharing the existing 4G spectrum with new 5G devices. The slot structure, subcarrier spacings, as well as flexibility in the timing and duration of transmission of different channels enable the network to schedule a mix of 5G users and 4G legacy devices in the same spectrum dynamically. 5G NR users can be scheduled alongside 4G users on the same spectrum, such that 5G transmissions do not interfere with any of the pre-configured LTE transmissions such as CRS, CSI-RS, PSS/SSS/PBCH, etc. This flexibility in 5G NR allows for a gradual re-farming of the bands where LTE is currently deployed for 5G, enables the early deployment of the 5G using the existing 4G bands, and, depending on implementation choices, may even allow reuse of existing base station site equipment with only a software upgrade required for enabling 5G.

Coexist: One of the design principles of 5G NR has been the ability to coexist within the LTE physical layer structure to allow for sharing existing 4G spectrum with new 5G devices.

5G NR UEs can simultaneously communicate over 5G-exclusive spectrum as well as spectrum shared with 4G users. There are two common scenarios worth highlighting in this regard. The first scenario is dual connectivity with a 4G anchor carrier enabling wide area coverage and mobility alongside one or more wide bandwidth 5G carriers providing high data-rate communication. The other is the use of 4G Uplink spectrum to serve as a 'Supplementary Uplink'. In cell edge conditions, the uplink may be power-limited and unable to take advantage of the wider bandwidths available in 5G spectrum, especially at higher frequencies. In such cases, uplink transmission over lower frequency spectrum may provide more reliable communication. Under such conditions, downlink communication occurs over 5G spectrum, while uplink transmissions occur on the lower frequency Supplementary Uplink carrier.

4 ... and Beyond

5G evolution continues beyond Release 15, enabling new services and applications, supporting new deployment scenarios, and unleashing new spectrum bands. The technology foundation set in Release 15 is also continually enhanced to provide a better latency, spectral efficiency, coverage, and power consumption in different sets of applications.

5G NR has already opened up vast swaths of spectrum for cellular communications. It continues to evolve further to enable deployments in additional frequency ranges.

– Operation of 5G NR in unlicensed bands is planned. Similar to the introduction of Licence Assisted Access (LAA) in LTE, this development opens up new spectrum for 5G devices to use on the downlink as well as uplink.
– Extending the deployment of 5G NR to spectral bands higher than 52.6 GHz is planned. The short-term focus in this area would be extension of support to 71 GHz and aim to rely heavily on the scalable numerology and flexible waveform and beam concepts already defined in 5G to support these spectrum bands.

New application domains for 5G NR are also under consideration.

– V2X communication (Vehicle-to-Vehicle, Vehicle-to-Infrastructure, and Vehicle-to-Pedestrian) is an emerging use case with applications in road safety and improved transportation efficiency. LTE-based C-V2X has already been developed. An extension to 5G NR is under consideration.
– The Internet of things spans a wide variety of devices with vastly different tiers of requirements. While 5G NR in its current form already enables a number of these tiers, low-cost and low-power devices deserve a special attention. To enable the support of a far higher density of low-cost and low-power devices (up to a million device per square kilometer), additional innovation is under way.
– Extended reality (XR) experience and applications such as virtual reality and augmented reality can benefit from the increased data throughputs and lower latency provided by 5G. Performance of the existing 5G networks for different use cases of XR and required enhancements is a study item in the near term.

Undoubtedly, as 5G deployments grow and new applications emerge, further evolution will occur to meet the demands of different use cases.

5 Conclusion

The development of 5G NR opens up yet another exciting chapter in cellular communication. Standing on the shoulders of the previous generations of wireless, 5G NR has been launched with ambitious goals to apply wireless communication to new vertical domains and applications, and to make use of vast amounts of the wireless spectrum that have hitherto never been used for cellular communication. These new requirements have led to a number of guiding principles such as operation in higher frequency bands, wider bandwidths, lower latency, flexibility in configuration and deployment, and coexistence with LTE. In addition, 5G NR has considered future extensibility to be one of its guiding principles, suggesting that 5G NR will have a long lifespan with future improvements occurring within the current overall framework without causing disruption to the existing use cases. The next few years will be interesting times, as 5G deployments grow and new applications and use cases emerge.

Abbreviations

5G: Fifth generation of wireless cellular communication; NR: New radio; 3GPP: Third Generation Partnership Project; FR1: Frequency Range 1 (Frequencies lower than 6 GHz); FR2: Frequency Range 2 (Frequencies higher than 24 GHz); AMPS: Advanced mobile phone system; GSM: Global system for mobile communications; TDMA: Time division multiple access; CDMA: Code division multiple access; WCDMA: Wideband CDMA; HSDPA: High speed data packet access; OFDM: Orthogonal frequency division multiplexing; OFDMA: Orthogonal frequency division multiple access; FFT: Fast Fourier Transform; UE: User equipment (refers to mobile device); FDD: Frequency division duplex; TDD: Time division duplex; LTE: Long term evolution.Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 16 December 2019 Accepted: 28 March 2020
Published online: 16 May 2020

References

1. 3GPP TS 38.211, NR: Physical channels and modulation. https://www.3gpp.org/ftp/Specs/archive/38_series/38.211/
2. 3GPP TS 38.212, NR; Multiplexing and channel coding. https://www.3gpp.org/ftp/Specs/archive/38_series/38.212/
3. 3GPP TS 38.213, NR; Physical layer procedures for control. https://www.3gpp.org/ftp/Specs/archive/38_series/38.213/
4. 3GPP TS 38.214, NR; Physical layer procedures for data. https://www.3gpp.org/ftp/Specs/archive/38_series/38.214/
5. 3GPP TS 38.215, NR: Physical layer measurements. https://www.3gpp.org/ftp/Specs/archive/38_series/38.215/

6. Dahlman E, Parkvall S, Skold J (2018) 5G NR The next generation wireless access technology. Elsevier, New York
7. Rappaport T, Heath RW Jr, Daniels RC, Murdock JN (2014) Millimeter wave wireless communications. Prentice Hall, Upper Saddle River
8. Dahlman E, Parkvall S, Skold J (2014) 4G: LTE/LTE-advanced for mobile broadband. Academic Press, Cambridge
9. Holma H, Toskala A (2000) Radio Access for Third Generation Mobile Communications. Wiley, WCDMA for UMTS, New York
10. Bender P, Black P, Grob M, Padovani R, Sindhushyana N, Viterbi A (2000) CDMA/HDR: a bandwidth efficient high speed wireless data service for nomadic users. IEEE Commun Mag 38(7):70–77
11. Viterbi AJ (1995) Principles of spread spectrum communication, wireless communication series, CDMA. Addison-Wesley, Boston
12. Subrahmanya P, Sundaresan R, Shiu DS (2004) An overview of high-speed packet data transport in CDMA systems. IETE Tech Rev 21(5):305–315
13. Mouly M, Pautet MB (1992) The GSM system for mobile communications. Telecom Publishing, London
14. Heegard C, Wicker SB (1999) Turbo coding. Springer, Berlin
15. Berrou C, Glavieux A, Thitimajshima P (1993) Near Shannon limit error-correcting coding and decoding: turbo-codes. In: Proceedings of ICC '93—IEEE international conference on communications, vol 38(7). pp 1064–1070
16. Robert GG (1963) The low density parity check codes. MIT Press, Cambridge
17. Arikan E, Polarization C (2009) A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. IEEE Trans Inf Theory 55(7):3051–73
18. Myung HG, Lim J, Goodman DJ (2006) Single carrier FDMA for uplink wireless transmission. IEEE Veh Technol Mag 1(3):30–38
19. Chang RC (1966) Synthesis of band-limited orthogonal signals for multichannel data transmission. Bell Syst Tech J 45(10):1775–1796
20. Paulraj A, Nabar R, Gore D (2003) Introduction to space-time communications. Cambridge University Press, Cambridge
21. Comroe R, Costello D (1984) ARQ schemes for data transmission in mobile radio systems. IEEE J Sel Areas Commun 2(4):472–481

Parvathanathan Subrahmanya is a wireless communications engineer and researcher in Silicon Valley. He has a B.Tech in Electronics and Communication Engineering from the Indian Institute of Technology, Madras and a Ph.D. in Electrical Engineering from Cornell University.

Amir Farajidana is a wireless communications engineer and researcher in Silicon Valley. He has a B.Sc. in Electrical Engineering from Sharif University of Technology and a Ph.D. in Electrical Engineering from California Institute of Technology.