CHAPTER 3

New Radio Access Physical Layer Aspects (Part 1)

This chapter describes the theoretical and practical aspects of physical layer protocols and functional processing in 3GPP new radio. As shown in Fig. 3.1, the physical layer is the lowest protocol layer in baseband signal processing that interfaces with the digital and the analog radio frontends and the physical media (in this case air interface) through which the signal is transmitted and received. The physical layer further interfaces with the medium access control (MAC) sublayer and receives MAC PDUs and processes the transport blocks through channel coding, rate matching, interleaving/scrambling, baseband modulation, layer mapping for multi-antenna transmission, digital precoding, resource element mapping, orthogonal frequency division multiplexing (OFDM) modulation, and antenna mapping. The choice of appropriate modulation and coding scheme as well as multi-antenna transmission mode is critical to achieve the desired reliability/robustness (coverage) and system/user throughput in mobile communications. Typical mobile radio channels tend to be dispersive and time variant and exhibit severe Doppler effects, multipath delay variation, intra-cell and inter-cell interference, and fading.

A good and robust design of the physical layer ensures that the system can robustly operate and overcome the above deleterious effects and can provide the maximum throughput and lowest latency under various operating conditions. Chapters 3 and 4 on physical layer in this book are dedicated to systematic design of physical layer protocols and functional blocks of 5G systems, the theoretical background on physical layer procedures, and performance evaluation of physical layer components. The theoretical background is provided to make the chapter self-contained and to ensure that the reader understands the underlying theory governing the operation of various functional blocks and procedures. While the focus is mainly on the techniques that were incorporated in the design of 3GPP NR physical layer, the author has attempted to take a more generic and systematic approach to the design of physical layer for the IMT-2020 wireless systems so that the reader can understand and apply the learnings to the design and implementation of any OFDM-based physical layer irrespective of the radio access technology.

5G NR. DOI: https://doi.org/10.1016/B978-0-08-102267-2.00003-8

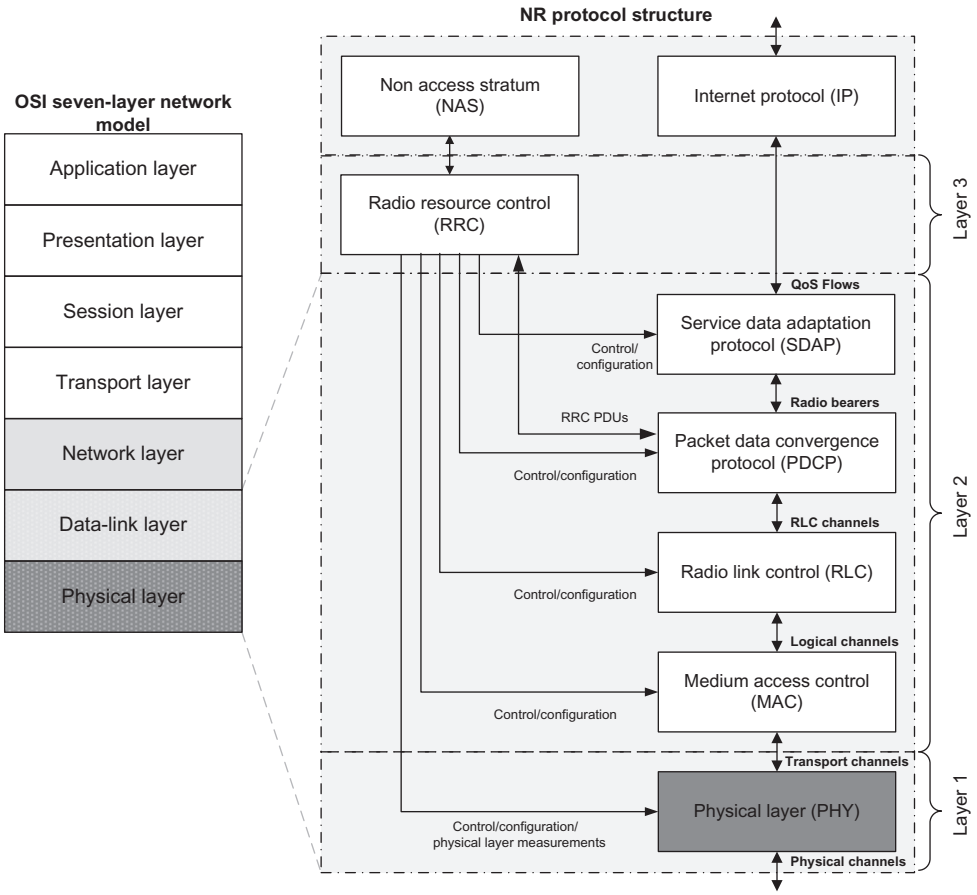


Figure 3.1
The physical layer in NR protocol stack [19].

In this chapter, we start with the study of the fundamental concepts and common features/functions in downlink and uplink of the new radio, which include review of the characteristics of wireless channels in sub-6 GHz and mmWave frequency regions as well as the analysis of two-dimensional (2D) and three-dimensional (3D) channel models and propagation effects. We will then begin our top-down approach to physical layer protocols starting with waveforms, orthogonal and non-orthogonal multiple-access schemes, and duplex schemes as well as the operating frequencies of the new radio. The frame structure, OFDM numerologies, time-frequency resources, and resource allocation techniques will be discussed and analyzed from theoretical and practical point of views.

3.1 Channel Models and Propagation Characteristics

3.1.1 Characteristics of Wireless Channels

In a wireless communication system, a signal can travel from the transmitter to the receiver over multiple paths. This phenomenon is referred to as multipath propagation where signal attenuation varies on different paths. This effect also known as multipath fading can cause stochastic fluctuations in the received signal's magnitude, phase, and angle of arrival (AoA). The propagation over different paths is caused by scattering, reflection, diffraction, and refraction of the radio waves by static and moving objects as well as the transmission medium. It is obvious that different propagation mechanisms result in different channel and path loss models. As a result of wave propagation over multipath fading channels, the radio signal is attenuated due to mean path loss as well as macroscopic and microscopic fading.

A detailed channel model for the frequency range (FR) from 6 to 100 GHz was developed in 3GPP [24]. It is applicable to bandwidth up to 10% of the carrier frequency (with a limit of 2 GHz), which accounts for the mobility of one of the two terminals (i.e., in a typical cellular network, the base station is fixed and the user terminal moves). It further provides several optional features that can be plugged in to the basic model, in order to simulate spatial consistency (i.e., the radio environment conditions of nearby users are correlated), blockage, and oxygen absorption. This model supports different IMT-2020 test environments including urban microcell (UMi), urban macrocell (UMa), rural macrocell (RMa), indoor hotspot (InH), and outdoor to indoor, which must be chosen when setting the simulation parameters [4,5,8,24,25].

3.1.1.1 Path Loss Models

In ideal free-space propagation model, the attenuation of RF signal energy between the transmitter and receiver follows inverse-square law. The received power expressed as a function of the transmitted power is attenuated proportional to the inverse of $L_s(d)$, which is called free-space path loss. When the receiving antenna is isotropic, the received signal power can be expressed as follows [32,33]:

$$P_{RX} = P_{TX} G_{TX} G_{RX} \left(\frac{\lambda}{4\pi d} \right)^2 = \frac{P_{TX} G_{TX} G_{RX}}{L_s(d)}$$

where P_{TX} and P_{RX} denote the transmitted and the received signal power, G_{TX} and G_{RX} denote the transmitting and receiving antenna gains, d is the distance between the transmitter and the receiver, and λ is the wavelength of the RF signal. In mmWave bands, the smaller wavelength translates into smaller captured energy at the receive antenna. For example, in the frequency range 3−30 GHz, an additional 20 dB path loss is added due to effective

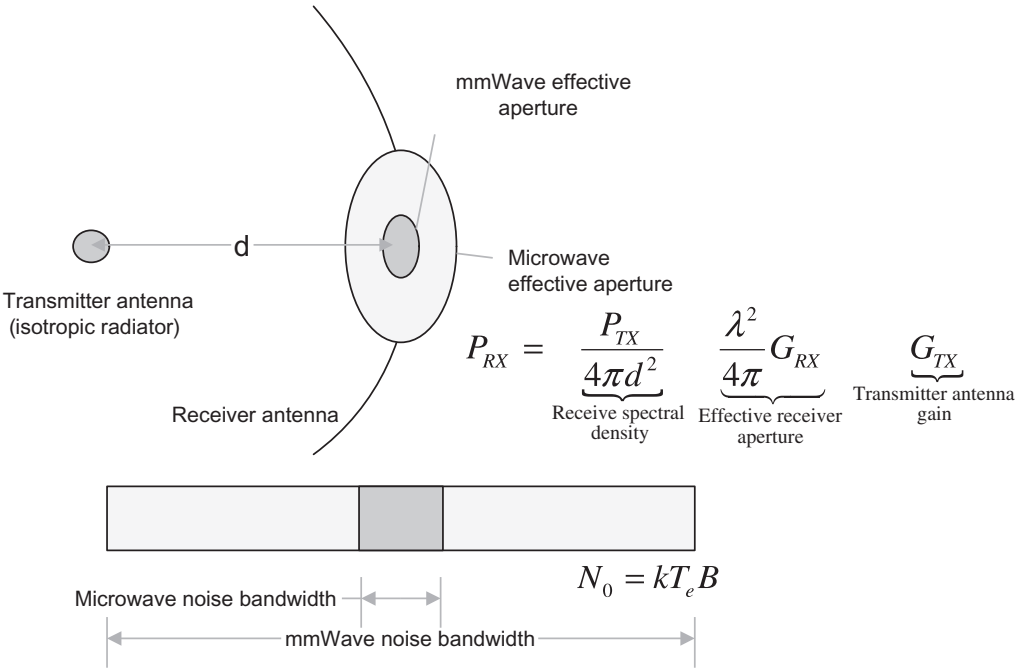


Figure 3.2
Definition of gain and aperture in mmWave [32].

aperture reduction. On the other hand, larger bandwidths in mmWave mean higher noise power and lower signal-to-noise ratios (SNRs), for example, from 50 to 500 MHz bandwidth, the noise power is increased by 10 dB (see Fig. 3.2).

Free-space conditions require a direct line of sight (LoS) between the two antennas involved. Consequently, no obstacles must exist in the path between the antennas at both ends. Furthermore, in order to avoid the majority of effects caused by superposition of direct and reflected signals, it is necessary that the first Fresnel zone[1] is completely free of obstacles. The first Fresnel ellipsoid is defined as a rotational ellipsoid with the two antennas at its focal points. Within this ellipsoid, the phase difference between two potential paths is less than half a wavelength. The radius b at the center of the ellipsoid can be calculated as $b = 8.66\sqrt{d/f}$ where b is the radius in meters, d is the distance between the receiver and the transmitter in kilometers, and f is the frequency in GHz [26,27,32].

[1] A Fresnel zone is one of a series of confocal prolate ellipsoidal regions of space between and around a transmitting antenna and a receiving antenna system. The regions are used to understand and compute the strength of waves (such as sound or radio waves) propagating between a transmitter and a receiver, as well as to predict whether obstructions near the line joining the transmitter and receiver will cause significant interference.

It must be noted that all antenna properties and attributes that we consider in this section assumes far-field patterns. The distance from an antenna, where far-field conditions are met, depends on the dimensions of the antenna relative to the wavelength. For smaller antennas, the wave fronts radiated from the antenna become almost parallel at much closer distance compared to electrically large antennas. A good approximation for small antennas is that far-field conditions are reached at $r \approx 2\lambda$. For larger antennas, that is, reflector antennas or array antennas where the dimensions of the antenna L are significantly larger compared to the wavelength $L \gg \lambda$, the far-field distance is approximated as $r \approx 2L^2/\lambda$ [26,33].

Macroscopic fading is caused by shadowing effects of buildings and natural obstructions and is modeled by the local mean of a fast fading signal. The mean path loss $\overline{L}_p(d)$ as a function of distance d between the transmitter and receiver is proportional to an nth power of d relative to a reference distance d_0. In logarithmic scale, it can be expressed as follows:

$$\overline{L}_p(d) = L_s(d_0) + 10n \log\left(\frac{d}{d_0}\right) \quad \text{(dB)}$$

The reference distance d_0 corresponds to a point located in the far-field of the antenna typically 1 km for large cells, 100 m for microcells, and 1 m for indoor channels. In the above equation, $\overline{L}_p(d)$ is the mean path loss which is typically $10n$ dB per decade attenuation for $d \gg d_0$. The value of n depends on the frequency, antenna heights, and propagation environment that is equal to 2 in free space. The studies show that the path loss $L_p(d)$ is a random variable with log-normal distribution about the mean path loss $\overline{L}_p(d)$. Let $X \sim \mathbf{N}(0, \sigma^2)$ denote a zero-mean Gaussian random variable with standard deviation σ when measured in decibels, then

$$L_p(d) = L_s(d_0) + 10n \log\left(\frac{d}{d_0}\right) + X \quad \text{(dB)}$$

The value of X is often derived empirically based on measurements. A typical value for σ is 8 dB. The parameters that statistically describe path loss due to large-scale fading (macroscopic fading) for an arbitrary location with a specific transmitting-receiving antenna separation include the reference distance d_0, the path loss exponent n, and the standard deviation σ of X [33].

Microscopic fading refers to the rapid fluctuations of the received signal in time and frequency and is caused by scattering objects between the transmitting and receiving antennas. When the received RF signal is a superposition of independent scattered components plus an LoS component, the envelope of the received signal $r(t)$ has a Rician Probability Distribution Function (PDF) that is referred to as Rician fading. As the

magnitude of the LoS component approaches zero, the Rician PDF approaches a Rayleigh PDF. Thus

$$f(r) = \frac{r(K+1)}{\sigma^2}\exp\left[-K - \frac{(K+1)r^2}{2\sigma^2}\right] I_0\left(\frac{2r}{\sigma}\sqrt{K\frac{(K+1)}{2}}\right), \quad r \geq 0$$

where K and $I_0(r)$ denote the Rician factor and zero-order modified Bessel function of the first kind.[2] In the absence of LoS path ($K = 0$), the Rician PDF reduces to Rayleigh distribution.

One of the challenges of mobile communications in the higher frequency bands for outdoor access has been to overcome the difficulties in highly varying propagation conditions. Understanding the propagation conditions will be critical to designing an appropriate air interface and determining the type of hardware (particularly the array size) needed for reliable communications. Extensive measurements over a wide range of frequencies were performed by a large number of academic institutions and the industry. Since maintaining link budgets at higher frequencies are challenging, there are few measurements at larger distances for 5G deployment scenarios of interest. Based on the results of the measurements, some important observations were made that helped development of the new channel models. The most notable signal degradation is due to the higher path loss of the bands above 6 GHz relative to sub-6 GHz bands. The additional path loss as result of increasing frequency need to be compensated by some means such as larger antenna array sizes with higher array gains and MIMO schemes to ensure sufficient and robust links. Due to large variation of propagation characteristics in bands above 6 GHz, the propagation characteristics of different frequency bands were independently investigated. The combined effects of all contributors to propagation loss can be expressed via the path loss exponent. In UMi test environment, the LoS path loss in the bands of interest appears to closely follow the free-space path loss model. In lower bands, a higher path loss exponent was observed in NLoS scenarios. The shadow fading in the measurements appears to be similar to lower frequency bands, while ray-tracing results show a much higher shadow fading (> 10 dB) than measurements, due to the larger dynamic range allowed in some ray-tracing experiments. In sub-6 GHz NLoS scenarios, the root mean square (RMS) delay spread is typically modeled in the range of 50−500 ns, the RMS azimuth angle spread of departure (from the access point) at around 10−30 degrees, and the RMS azimuth angle spread of arrival (at the UE) at around 50−80 degrees [7,8]. There are measurements of the delay spread above 6 GHz which indicate somewhat smaller ranges as the frequency increases, and some measurements show the millimeter wave omnidirectional channel to be highly directional in nature.

[2] A modified Bessel function of the first kind is a function $I_n(x)$ which is one of the solutions to the modified Bessel differential equation and is closely related to the Bessel function of the first kind $J_n(x)$. The modified Bessel function of the first kind $I_n(z)$ can be defined by the contour integral $I_n(z) = \oint e^{z/2(t+1)/t} t^{-n-1} dt$ where the contour encloses the origin and is traversed in a counterclockwise direction. In terms of $J_n(x)$, $I_n(x) \triangleq j^{-n} J_n(jx) = e^{-jn\pi/2} J_n(xe^{j\pi/2})$.

In UMa test environments, the behavior of LoS path loss is similar to free-space path loss and the NLoS path loss behavior appears not following a certain model over a wide range of frequencies. It was observed that the rate at which the path loss increased with frequency was not linear, as the rate is higher in the lower parts of the spectrum, which can be possibly explained as due to diffraction effect that is frequency-dependent and a more dominating component in lower frequencies. However, in higher frequencies, reflections and scattering are relatively the predominant components. From preliminary ray-tracing studies, the channel spreads in terms of delay and angle appear to be weakly dependent on the frequency and are generally 2−5 times smaller than the values reported in [7]. The cross-polar scattering in the ray-tracing results tends to increase with increasing frequency due to diffuse scattering.[3]

In InH deployment scenarios, under LoS conditions, multiple reflections from walls, floor, and ceiling could give rise to wave-guiding effect. The measurements conducted in office scenarios suggest that path loss exponent, based on a 1 m free-space reference distance is typically below 2, leading to relatively less path loss than predicted by the free-space path loss formula. The strength of the wave-guiding effect is variable and the path loss exponent appears to increase slightly with increasing frequency, possibly due to the relation between the wavelength and surface roughness. Measurements of the small-scale channel properties such as angular spread (AS) and delay spread have shown similarities between channels over a very wide frequency range, where the results suggest that the main multipath components are present at all frequencies with some small variations in magnitudes. Recent studies have shown that polarization discrimination ranges between 15 and 25 dB for indoor mmWave channels with greater polarization discrimination at 73 GHz than at 28 GHz [1−3].

Cross-polarized signal components are generated by reflection and diffraction. It is widely known that the fading correlation characteristic between orthogonally polarized antennas has a very low correlation coefficient. Polarization diversity techniques and MIMO systems with orthogonally polarized antennas are developed that employ this fading characteristic. Employing the polarization diversity technique is one solution to improving the received power, and the effect of the technique is heavily dependent on the cross-polarization discrimination (XPD) ratio characteristic. Moreover, the channel capacity can be improved by appropriately using the cross-polarization components in MIMO systems. Thus the communication quality can be improved by effectively using the information regarding the cross-polarized waves in a wireless system. XPD is an important characteristic, particularly in

[3] Diffuse scattering is a form of scattering that arises from deviation of material structure from that of a perfectly regular lattice, which appears in experimental data as scattering spread over a wide $q-$ range (diffuse). Diffuse scattering is generally difficult to quantify and is closely related to Bragg diffraction, which occurs when scattering amplitudes add constructively. The defect in a crystal lattice results in reduction of the amplitude of the Bragg peak.

dual-polarized systems, where cross-talk between polarizations can prevent the system to achieve its quality objectives. Radios can use cross-polarization interference cancellation (XPIC) to isolate polarizations and compensate for any link or propagation induced coupling. However, good antenna polarization is important to allow the XPICs maximum flexibility to compensate for these dynamic variations.

The following equations provide one of the path loss models for UMa test environment for LoS and NLoS scenarios in the frequency range of 0.5−100 GHz. These models were developed based on measurement results conducted by independent academic and industry organizations and published in the literature [7,24,34−36].

- LoS path loss for 0.5 GHz $\leq f_c \leq$ 100 GHz

$$PL_{UMa-LoS} = \begin{cases} PL_1 & 10 \text{ m} \leq d_{2D} \leq d'_{BP} \\ PL_2 & d'_{BP} \leq d_{2D} \leq 5 \text{ km} \end{cases}$$

$$PL_1 = 32.4 + 20\log_{10}(d_{3D}) + 20\log_{10}(f_c)$$

$$PL_2 = 32.4 + 40\log_{10}(d_{3D}) + 20\log_{10}(f_c) - 10\log_{10}\left[\left(d'_{BP}\right)^2 + \left(h_{gNB} - h_{UE}\right)^2\right]$$

$$h_{gNB} = 25 \text{ m}, \quad 1.5 \text{ m} \leq h_{UE} \leq 22.5 \text{ m}, \quad \sigma_{SF} = 4 \text{ dB}$$

- NLoS path loss for 6 GHz $\leq f_c \leq$ 100 GHz

$$PL_{UMa-NLoS} = \max\left(PL_{UMa-LoS}, PL'_{UMa-NLoS}\right)$$

$$10 \text{ m} \leq d_{2D} \leq 5 \text{ km}$$

$$PL'_{UMa-NLoS} = 13.54 + 39.08\log_{10}(d_{3D}) + 20\log_{10}(f_c) - 0.6(h_{UE} - 1.5)$$

$$1.5 \text{ m} \leq h_{UE} \leq 22.5 \text{ m}, h_{gNB} = 25 \text{ m}, \sigma_{SF} = 6$$

3.1.1.2 Delay Spread

Time-varying fading due to scattering objects or transmitter/receiver motion results in Doppler spread. The time spreading effect of small-scale or microscopic fading is manifested in the time domain as multipath delay spread and in the frequency domain as channel coherence bandwidth. Similarly, the time variation of the channel is characterized in the time domain as channel coherence time and in the frequency domain as Doppler spread. In a fading channel, the relationship between maximum excess delay time τ_m and symbol time τ_s can be viewed in terms of two different degradation effects, that is, frequency-selective fading and frequency non-selective or flat fading. A channel is said to exhibit frequency-selective fading if $\tau_m > \tau_s$. This condition occurs whenever the received multipath components of a symbol extend beyond the symbol's time duration. Such multipath dispersion of the signal results in inter-symbol interference (ISI) distortion. In the case of frequency-selective fading, mitigating the distortion is possible because many of the multipath components are separable by the receiver. A channel is said to exhibit frequency non-selective or

flat fading if $\tau_m < \tau_s$. In this case, all multipath components of a symbol arrive at the receiver within the symbol time duration; therefore, the components are not resolvable. In this case, there is no channel-induced ISI distortion, since the signal time spreading does not result in significant overlap among adjacent received symbols. There is still performance degradation because the irresolvable phasor components can add up destructively to yield a substantial reduction in SNR. Also, signals that are classified as exhibiting flat fading can sometimes experience frequency-selective distortion [33].

Fig. 3.3 illustrates multipath-intensity profile $\Lambda(\tau)$ versus delay τ where the term delay refers to the excess delay. It represents the signal's propagation delay that exceeds the delay of the first signal component arrival at the receiver. For a typical wireless communication

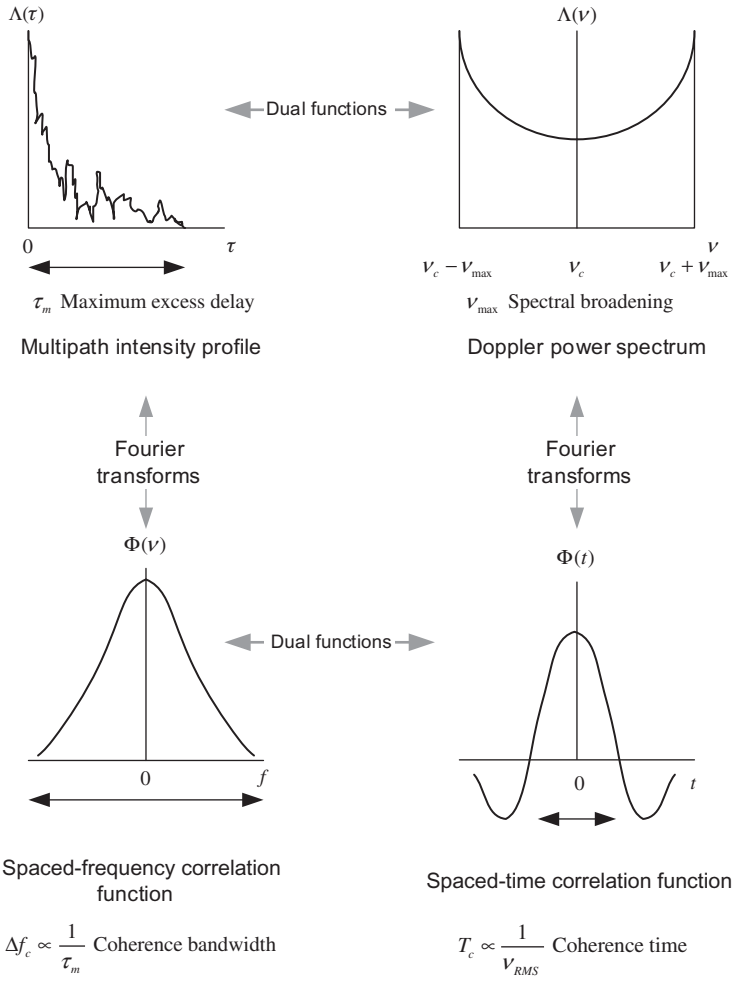


Figure 3.3
Illustration of the duality principle in time and frequency domains [33].

channel, the received signal usually consists of several discrete multipath components. The received signals are composed of a continuum of multipath components in some channels such as the tropospheric channel. In order to perform measurements of the multipath-intensity profile, a wideband signal, that is, a unit impulse or Dirac delta function, is used. For a single transmitted impulse, the time τ_m between the first and last received component is defined as the maximum excess delay during which the multipath signal power typically falls to some level 10−20 dB below that of the strongest component. Note that for an ideal system with zero excess delay, the function $\Lambda(\tau)$ would consist of an ideal impulse with weight equal to the total average received signal power. In the literature, the Fourier transform of $\Lambda(\tau)$ is referred to as spaced-frequency correlation function $\Phi(\upsilon)$. The spaced-frequency correlation function $\Phi(\upsilon)$ represents the channel's response to a pair of sinusoidal signals separated in frequency by υ. The coherence bandwidth Δf_c is a measure of the frequency range over which spectral components have a strong likelihood of amplitude correlation. In other words, a signal's spectral components over this range are affected by the channel in a similar manner. Note that Δf_c and τ_m are inversely proportional $\left(\Delta f_c \propto 1/\tau_m\right)$. The maximum excess delay τ_m is not the best indicator of how a given wireless system will perform over a communication channel because different channels with the same value of τ_m can exhibit different variations of signal intensity over the delay span. The delay spread is often characterized in terms of the RMS delay spread τ_{RMS} in which the average multipath delay $\overline{\tau}$ is calculated as follows:

$$\tau_{RMS} = \sqrt{\frac{\int_0^{\tau_m} (\tau - \overline{\tau})^2 \Lambda(\tau) d\tau}{\int_0^{\tau_m} \Lambda(\tau) d\tau}}, \quad \overline{\tau} = \frac{\int_0^{\tau_m} \tau \Lambda(\tau) d\tau}{\int_0^{\tau_m} \Lambda(\tau) d\tau}$$

An exact relationship between coherence bandwidth and delay spread does not exist and must be derived from signal analysis of actual signal dispersion measurements in specific channels. If coherence bandwidth is defined as the frequency interval over which the channel's complex-valued frequency transfer function has a correlation of at least 0.9, the coherence bandwidth is approximately $\Delta f_c \approx 1/(50\tau_{RMS})$. A common approximation of Δf_c corresponding to a frequency range over which the channel transfer function has a correlation of at least 0.5 is $\Delta f_c \approx 1/(5\tau_{\mathrm{RMS}})$.

A channel is said to exhibit frequency-selective effects, if $\Delta f_c < 1/\tau_s$ where the inverse symbol rate is approximately equal to the signal bandwidth W. In practice, W may differ from $1/\tau_s$ due to filtering or data modulation. Frequency-selective fading effects arise whenever a signal's spectral components are not equally affected by the channel. This occurs whenever $\Delta f_c < W$. Frequency non-selective or flat fading degradation occurs whenever $\Delta f_c > W$. Hence, all spectral components of the signal will be affected by the channel in a similar manner. Flat fading does not introduce channel-induced ISI distortion, but performance degradation can still be expected due to loss in SNR, whenever the signal

experiences fading. In order to avoid channel-induced ISI distortion, the channel is required to exhibit flat-fading by ensuring that $\Delta f_c > W$. Therefore, the channel coherence bandwidth Δf_c sets an upper limit on the transmission rate that can be used without incorporating an equalizer in the receiver [33].

3.1.1.3 Doppler Spread

Fig. 3.3 shows another function $\Phi(t)$ known as spaced-time correlation function, which is the autocorrelation function of the channel's response to a sinusoid. This function specifies the extent to which there is correlation between the channel's response to a sinusoid sent at time t_1 and the response to a similar sinusoid sent at time t_2, where $\Delta t = t_2 - t_1$. The coherence time is a measure of the expected time duration over which the channel's response is essentially invariant. To estimate $\Phi(t)$, a sinusoidal signal is transmitted through the channel and the autocorrelation function of the channel output is calculated. The function $\Phi(t)$ and the coherence time T_c provide information about the rate of fading channel variation. Note that for an ideal time-invariant channel, the channel's response would be highly correlated for all values of Δt and $\Phi(t)$ would be a constant function. If one ideally assumes uniformly distributed scattering around a mobile station with linearly polarized antennas, the Doppler power spectrum (i.e., the inverse Fourier transform of spaced-time correlation function $\Lambda(\nu)$) has a U-shaped distribution as shown in Fig. 3.3. In a time-varying fading channel, the channel response to a pure sinusoidal tone spreads over a finite frequency range $\nu_c - \nu_{\max} < \nu < \nu_c + \nu_{\max}$, where ν_c and $\nu_{\max}$ denote the frequency of the sinusoidal tone and the maximum Doppler spread, respectively. The RMS bandwidth of $\Lambda(\nu)$ is referred to as Doppler spread and is denoted by ν_{RMS} that can be estimated as follows:

$$\nu_{RMS} = \sqrt{\frac{\int_{\nu_c-\nu_{\max}}^{\nu_c+\nu_{\max}} (\nu-\overline{\nu})^2 \Lambda(\nu) d\nu}{\int_{\nu_c-\nu_{\max}}^{\nu_c+\nu_{\max}} \Lambda(\nu) d\nu}} \quad \overline{\nu} = \frac{\int_{\nu_c-\nu_{\max}}^{\nu_c+\nu_{\max}} \nu \Lambda(\nu) d\nu}{\int_{\nu_c-\nu_{\max}}^{\nu_c+\nu_{\max}} \Lambda(\nu) d\nu}$$

The coherence time is typically defined as the time lag for which the signal autocorrelation coefficient reduces to 0.7. The coherence time is inversely proportional to Doppler spread $T_c \approx 1/\nu_{RMS}$. A common approximation for the value of coherence time as a function of Doppler spread is $T_c = 0.423/\nu_{RMS}$. It can be observed that the functions on the right side of Fig. 3.3 are dual of the functions on the left side (duality principle).

3.1.1.4 Angular Spread

The angle spread refers to the spread in AoA of the multipath components at the receiver antenna array. At the transmitter, on the other hand, the angle of spread refers to the spread in the angle of departure (AoD) of the multipath components that leave the transmit

antennas. If the angle spectrum function $\Theta(\theta)$ denotes the average power as function of AoA, then the RMS angle spread can be estimated as follows:

$$\theta_{RMS} = \sqrt{\frac{\int_{-\pi}^{+\pi}(\theta-\overline{\theta})^2\Theta(\theta)d\theta}{\int_{-\pi}^{+\pi}\Theta(\theta)d\theta}} \quad \overline{\theta} = \frac{\int_{-\pi}^{+\pi}\Theta(\theta)\theta d\theta}{\int_{-\pi}^{+\pi}\Theta(\theta)d\theta}$$

The angle spread causes space-selective fading, which manifests itself as variation of signal amplitude according to the location of antennas. The space-selective fading is characterized by the coherence distance D_c which is the spatial separation for which the autocorrelation coefficient of the spatial fading reduces to 0.7. The coherence distance is inversely proportional to the angle spread $D_c \propto 1/\theta_{RMS}$. In Fig. 3.3 a duality between multipath-intensity function $\Lambda(\tau)$ and Doppler power spectrum $\Lambda(\nu)$ is shown, which means that the two functions exhibit similar behavior across time domain and frequency domain. As the $\Lambda(\tau)$ function identifies expected power of the received signal as a function of delay, $\Lambda(\nu)$ identifies expected power of the received signal as a function of frequency. Similarly, spaced-frequency correlation function $\Phi(f)$ and spaced-time correlation function $\Phi(t)$ are dual functions. It implies that as $\Phi(f)$ represents channel correlation in frequency, $\Phi(t)$ corresponds to channel correlation function in time in a similar manner [33].

The AS in radians can be expressed based on the circular standard deviation in directional statistics using the following expression [24,25]:

$$\text{AS} = \sqrt{-2\ \log\left(\left|\frac{\sum_{n=1}^{N}\sum_{m=1}^{M}(e^{j\varphi_{nm}}P_{nm})}{\sum_{n=1}^{N}\sum_{m=1}^{M}P_{nm}}\right|\right)}$$

where P_{nm} denotes the power of the mth subpath of the nth path and φ_{nm} is the subpaths angle (either AoA, AoD, elevation angle of arrival [EoA], elevation angle of departure [EoD]). In order to model large signal bandwidths and large antenna arrays, the channel models have been specified with sufficiently high resolution in the delay and angular domains. There are two important aspects related to large antenna arrays. One is the very large size of the antenna array and the other is the large number of antenna elements. These features require high angular resolution in channel modeling, which means more accurate modeling of AoA/AoD, and possibly higher number of multipath components.

3.1.1.5 Blockage

The blockage model describes a phenomenon where the stationary or moving objects standing between the transmitter and receiver dramatically change the channel characteristics and in some cases may block the signal, especially in high-frequency bands, since the signal in mmWave does not effectively penetrate or diffract around human bodies and other objects. Shadowing by these objects is an important factor in the link budget calculations and the time variation of the channel, and such dynamic blocking may be important to capture in

evaluations of technologies that include beam-finding and beam-tracking capabilities. The effect of the blockage is considered not only on the total received power, but also on the angle or power of multipath due to different size, location, and direction of the blocker. There are two categories of blockage: (1) dynamic blockage and (2) geometry-induced blockage. Dynamic blockage is caused by the moving objects in the communication environment. The effect is additional transient loss on the paths that intercept the moving objects. Geometry-induced blockage, on the other hand, is a static property of the environment. It is caused by objects in the map environment that block the signal paths. The propagation channels in geometry-induced blockage locations are dominated by diffraction and sometimes by diffuse scattering. The effect is an additional loss beyond the normal path loss and shadow fading. Compared to shadow fading caused by reflections, diffraction-dominated shadow fading may have different statistics (e.g., different mean, variance, and coherence distance) [33]. Radio waves are attenuated by foliage, and this effect increases with frequency. The main propagation phenomena involved are attenuation of the radiation through the foliage, diffraction above/below and sideways around the canopy, and diffuse scattering by the leaves. The vegetation effects are captured implicitly in the path loss equations.

A stochastic method for capturing human and vehicular blocking in mmWave frequency regions can be used (among other methods) to model the blockage effect. In this case, the number of blockers must be first determined. For this purpose, multiple 2D angular blocking regions, in terms of center angle, azimuth, and elevation angular span are generated around the UE. There is one self-blocking region and $K = 4$ non-self-blocking regions, where K may be changed for certain scenarios such as higher blocker density. Note that the self-blocking component of the model is important in capturing the effects of human body blocking. In the next step, the size and location of each blocker must be generated. For self-blocking, the blocking region in the UE local coordinate system[4] is defined in terms of elevation and azimuth angles $(\theta'_{sb}, \varphi'_{sb})$ and azimuth and elevation angular span (x_{sb}, y_{sb}) [4,5].

$$\left\{(\theta', \varphi') \middle| \left(\theta'_{sb} - \frac{y_{sb}}{2} \le \theta' \le \theta'_{sb} + \frac{y_{sb}}{2}, \quad \varphi'_{sb} - \frac{x_{sb}}{2} \le \varphi' \le \varphi'_{sb} + \frac{x_{sb}}{2}\right)\right\}$$

where the parameters of the above equation are described in Table 3.1. For non-self-blocking $k = 1, 2, 3, 4$, the blocking region in global coordinate system is defined as follows [4,5]:

$$\left\{(\theta, \varphi) \middle| \left(\theta_k - \frac{y_k}{2} \le \theta \le \theta_k + \frac{y_k}{2}, \quad \varphi_k - \frac{x_k}{2} \le \varphi \le \varphi_k + \frac{x_k}{2}\right)\right\}$$

where d is the distance between the UE and the blocker; other parameters are given in Table 3.1.

[4] Global and local coordinate systems are used to locate geometric items in space. By default, a node coordinates are defined in the global Cartesian system.

Table 3.1: Blocking region parameters [4,5].

Self-Blocking Region Parameters					
Mode	φ'_{sb}	x_{sb}	θ'_{sb}	y_{sb}	
Portrait (degree)	260	120	100	80	
Landscape (degree)	40	160	110	75	
Blocking Region Parameters					
Blocker Index $k = 1, 2, 3, 4$	φ_k	x_k	θ_k	y_k	d
InH scenario	Uniform in [0°, 360°]	Uniform in [15°, 45°]	90°	Uniform in [5°,15°]	2 m
UMi_x, Uma_x, RMa_x scenarios	Uniform in [0°, 360°]	Uniform in [5°,15°]	90°	5°	10 m

The attenuation of each cluster due to self-blocking corresponding to the center angle pair $(\theta'_{sb}, \varphi'_{sb})$ $(\theta'_{sb}, \varphi'_{sb})$ is 30 dB provided that $\left|\varphi'_{AOA} - \varphi'_{sb}\right| < x_{sb}/2$ and $\left|\theta'_{ZOA} - \theta'_{sb}\right| < y_{sb}/2$; otherwise, the attenuation is 0 dB. Note that ZOA denotes the Zenith angle of arrival. The attenuation of each cluster due to the non-self-blocking regions $k = 1, 2, 3, 4$ is given as $L_{\text{dB}} = -20\ \log_{10}\left[1 - \left(F_{A_1} + F_{A_2}\right)\left(F_{Z_1} + F_{Z_2}\right)\right]$ provided that $|\varphi_{AOA} - \varphi_k| < x_k$ and $|\theta_{ZOA} - \theta_k| < y_k$; otherwise, the attenuation is 0 dB. The terms $F_{A_1|A_2|Z_1|Z_2}$ in the previous equation are defined as follows [4,5]:

$$F_{A_1|A_2|Z_1|Z_2} = \frac{1}{\pi}\tan^{-1}\left[\pm\frac{\pi}{2}\sqrt{\frac{\pi}{\lambda}d\left(\frac{1}{\cos(A_1|A_2|Z_1|Z_2)} - 1\right)}\right]$$

where $A_1 = \varphi_{AOA} - \left(\varphi_k + x_k/2\right)$, $A_2 = \varphi_{AOA} - \left(\varphi_k - x_k/2\right)$, $Z_1 = \theta_{ZOA} - \left(\theta_k + y_k/2\right)$, and $Z_2 = \theta_{ZOA} - \left(\theta_k - y_k/2\right)$. The center of the blocker is generated based on a uniformly distributed random variable, which is temporally and spatially consistent. The 2D autocorrelation function $R(\Delta_x, \Delta_t)$ can be described with sufficient accuracy by the exponential function $R(\Delta_x, \Delta_t) = \exp\left[-\left(|\Delta_x|/d_{corr} + |\Delta_t|/t_{corr}\right)\right]$ where d_{corr} denotes the spatial correlation distance or the random variable determining the center of the blocker and t_{corr} is the correlation time given as $t_{corr} = d_{corr}/v$, in which v is the speed of moving blocker [4,5].

3.1.1.6 Oxygen Absorption

The electromagnetic wave may be partially or totally attenuated by an absorbing medium due to atomic and molecular interactions. This gaseous absorption causes additional loss to the radio wave propagation. For frequencies around 60 GHz, additional loss due to oxygen absorption is applied to the cluster responses for different center frequency and bandwidth correspondingly. The additional loss $OL_n(f_c)$ for cluster n at center frequency f_c is given as follows [4]:

$$OL_n(f_c) = \frac{\alpha(f_c)[d_{3D} + c(\tau_n + \tau_\Delta)]}{1000}\ \text{[dB]}$$

Table 3.2: Oxygen attenuation $\alpha(f_c)$ as a function of frequency [4,5].

Frequency (GHz)	0–52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68–100
$\alpha(f_c)$(dB/km)	0	1	2.2	4	6.6	9.7	12.6	14.6	15	14.6	14.3	10.5	6.8	3.9	1.9	1	0

where $\alpha(f_c)$ denotes the frequency-dependent oxygen absorption loss in dB/km whose sample values at some frequencies are shown in Table 3.2, c is the speed of light in m/s, d_{3D} is the 3D distance in meters between the receive and transmit antennas, τ_n is the nth cluster delay in seconds, and $\tau_\Delta = 0$ in the LoS scenarios. For center frequencies not shown in Table 3.2, the frequency-dependent oxygen absorption loss $\alpha(f_c)$ is obtained from a linear interpolation of the values corresponding to the two adjacent frequencies [4,5].

For wideband channels, the time-domain channel response of each cluster (all rays within one cluster share common oxygen absorption loss) are transformed into frequency-domain channel response and the oxygen absorption loss is applied to the cluster's frequency-domain channel response for frequency $f_c + \Delta f$ within the channel bandwidth W. The oxygen loss $OL_n(f_c + \Delta f)$ for cluster n at frequency $f_c + \Delta f$, where $-W/2 \le \Delta f \le W/2$ is given as follows:

$$OL_n(f_c + \Delta f) = \frac{\alpha(f_c + \Delta f)[d_{3D} + c(\tau_n + \tau_\Delta)]}{1000} \text{ [dB]}$$

where $\alpha(f_c + \Delta f)$ is the oxygen absorption loss in dB/km at frequency $f_c + \Delta f$. The final frequency-domain channel response is obtained by the summation of frequency-domain channel responses of all clusters. Time-domain channel response is obtained by the reverse transform from the obtained frequency-domain channel response [4,5].

Measurements in mmWave frequency bands have shown that ground reflection in mmWave has significant effect which can produce a strong propagation path that superimposes with the direct LoS path and induces severe fading effects. When ground reflection is considered, the randomly generated shadow fading is largely replaced by deterministic fluctuations in terms of distance. As a result, the standard deviation of shadow fading, when ground reflection is considered, is set to 1 dB. The value of 1 dB was obtained via simulations in order to maintain a similar level of random channel fluctuations without ground reflection [4,5].

The mmWave channels are sparse, that is, they have few entries in the delay angle bins, although experimental verification of this may be limited due to the resolution of rotating horn antennas used for such measurements. However, a lower bound on the channel sparsity can still be established from existing measurements, and in many environments, the percentage of delay/angle bins with significant energy is rather low but not necessarily lower than at centimeter-wave frequencies.

Molecular oxygen absorption around 60 GHz is particularly high (~13 dB/km, depending on the altitude) but decreases rapidly away from the oxygen resonance frequency to below 1 dB/km. While these absorption values are considered high for macrocell links, cell densification has already reduced the required link distance to a substantially smaller range in urban areas, and the densification process will continue to reduce cell sizes. For a link distance of approximately 200 m, the path loss of a 60 GHz link in heavy rain condition is less than 3 dB. Fig. 3.4 illustrates the impact of rainfall and oxygen absorption throughout the mmWave transmission. A more serious issue than free-space loss for mmWave signals is their limited penetration through materials and limited diffraction. In the urban environment, coverage for large cells could be particularly challenging; however, cell densification can be used to achieve the ambitious 5G capacity goals.

Atmospheric effects such as oxygen and water vapor absorption as well as fog and precipitation can scale exponentially with the link distance. Limiting to distances below 1 km, attenuation caused by atmospheric gases can be neglected up to 50 GHz, as shown in Fig. 3.4. However, above 50 GHz, it becomes important to consider the oxygen absorption peak of approximately 13 dB/km at 60 GHz and the water vapor resonance peak at 183 GHz of approximately 29 dB/km for relative humidity of 44% under standard

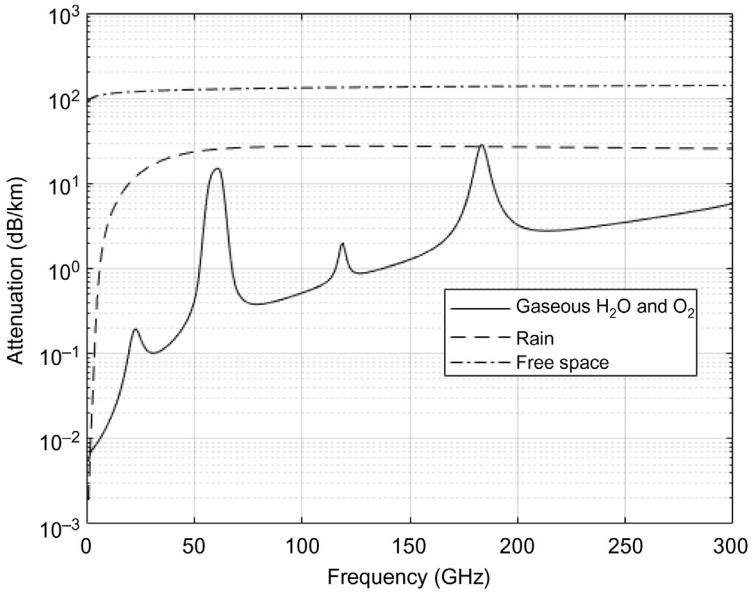


Figure 3.4
Comparison of gaseous $H_2O + O_2$, rain, and free-space attenuation (propagation distance 1 km, rain rate of 95 mm/h, and dry air pressure is 1013 hPa and the water vapor density is 7.5 g/m^3).

conditions. Note that neither fog nor rain is relevant for frequencies below 6 GHz. At frequencies above 80 GHz, dense fog related to a visibility of less than 70 m has a noticeable impact (>3 dB/km) and becomes severe above 200 GHz (>10 dB/km). Drizzle and steady rain are not a substantial issue for distances up to 1 km (3.0−4.4 dB/km above 70 GHz). However, as shown in Fig. 3.4, heavy rain attenuation increases up to 10−15 dB/km, and for downpours, up to 40 dB/km can be experienced. In summary, atmospheric effects, especially under bad weather conditions, are relevant for mmWave links over distances greater than 100 m and a crucial issue for longer distances of 1 km and farther [32].

3.1.1.7 LoS Path Loss Probability

It is shown in the literature that a height-dependent path loss for an indoor UE associated with a LoS condition can be modeled considering the dimensions of the building and the location of the UE inside the building. 3GPP has modeled the LoS path loss by using the 3D distance between the gNB and the UE along with the coefficients given by the ITU-R LoS path loss equations for 3D-UMa and 3D-UMi [4,24]. This provides a reasonable approximation to the more accurate models and can be determined without explicitly modeling the building dimensions. The ITU-R LoS path loss model assumes a two-ray model resulting in a path loss equation transitioning from a 22 dB/decade slope to a steeper slope at a break point depending on the environment height, which represents the height of a dominant reflection from the ground (or a moving platform) that can add constructively or destructively to the direct ray received at a UE located at the street level. In the 3D-UMa scenario, it is likely that such a dominant reflection path may come from the street level for indoor UEs associated with a type-1 LoS condition. Therefore, the environmental height is fixed at 1 m for a UE associated with a type-1 LoS condition. In the case of a UE associated with a type-2 LoS condition a dominant path can be likely created by reflection from the rooftop of a neighboring building. Note that a rooftop is at least 12 m in height; in this case the environmental height is randomly determined from a discrete uniform distribution between the UE height in meter and 12 [4,24].

For NLoS path loss modeling, which is the primary radio propagation mechanism in a 3D-UMa scenario, the dominant propagation paths experience multiple diffractions over rooftops followed by diffraction at the edge of the building. The path loss attenuation increases with the diffraction angle as a UE transitions from a high floor to a lower floor. In order to model this phenomenon, a linear height gain term given by $-\alpha(h-1.5)$ is introduced, where α in dB/m is the gain coefficient. A range of values between 0.6 and 1.5 were observed in different studies based on field measurements and ray-tracing simulations and a nominal value of 0.6 dB/m was chosen. In 3D-UMi test environments, the dominant propagation paths travel through and around the buildings. The UE may also receive a small

amount of energy from propagation above rooftops. In order to simplify the model, a linear height gain is also applied to the 3D-UMi NLoS path loss with 0.3 dB/m gain coefficient based on the results from multiple studies. In addition, in both 3D-UMa and 3D-UMi scenarios the NLoS path loss is lower bounded by the corresponding LoS path loss because the path loss in a NLoS environment is in principle larger than that in a LoS environments. Measurement results and ray-tracing data have indicated that the marginal distribution of the composite power angular spectrum in zenith statistically has a Laplacian distribution and its conditional distribution given a certain link distance and UE height can also be approximated by Laplacian distribution. To incorporate these observations, zenith angle of departure (ZoD) and zenith angle of arrival (ZoA) are modeled by inverse Laplacian functions. It is also observed that the zenith angle spread of departure (ZSD) decreases significantly as the UE moves further away from the gNB. An intuitive explanation is that the angle subtended by a fixed local ring of scatterers at the UE to the gNB decreases as the UE moves away from the gNB. The ZSD is also observed to slightly change as an indoor UE moves up to higher floors [34–36].

We use the concept of the LoS probability to distinguish between the LoS and NLoS links. A link of length d is LoS with probability $p_{LoS}(d)$. The LoS probability is a non-increasing function of the link length. The LoS probability is obtained based on the certain building models, that is, either we use stochastic models from random shape theory or we use site-specific maps from geographical information system database. In this section, the LoS condition is determined based on a map-based approach, that is, by considering the transmitter and receiver positions and whether any buildings or walls are blocking the direct path between them. The impact of in-between objects not represented in the map is modeled separately through shadowing or blocking path loss components. It is noteworthy that this LoS definition is frequency independent, due to the fact that only buildings and walls are considered in the definition. 3GPP and ITU-R define the UMa LoS probability as follows:

$$p_{LoS}(d) = \min\left(\frac{d_1}{d_{2D}}, 1\right)\left(1 - e^{-d_{2D}/d_2}\right) + e^{-d_{2D}/d_2}$$

where d_{2D} is the 2D distance in meters and d_1 and d_2 can be optimized to fit a set of measurement data in the test environments/scenarios under consideration. For UMi test environments, it was observed that the above LoS probability is sufficient for frequencies above 6 GHz. The fitted d_1/d_2 model provides more consistency with measured data and the error between the measured data and the 3GPP LoS probability model over all distances are small. Note that the 3GPP UMi LoS probability model is not a function of UE height unlike the UMa LoS probability model.

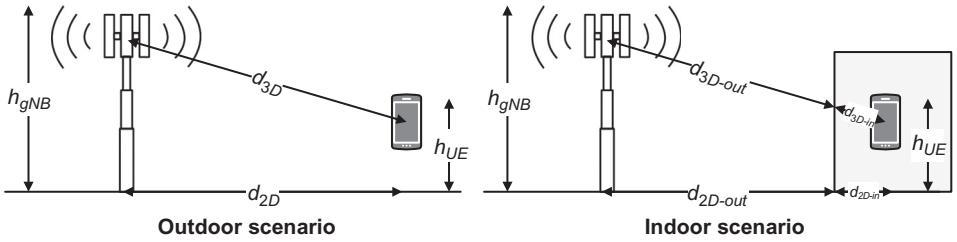


Figure 3.5
Definition of 2D and 3D distances in outdoor/indoor environments [4].

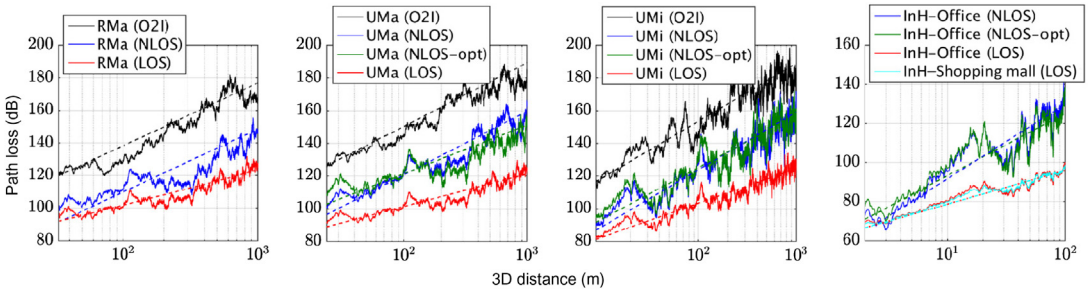


Figure 3.6
Path loss with (*solid line*) and without (*dashed line*) shadowing for various scenarios [37].

In the path loss models we often refer to 2D and 3D distances. Those distances are defined as follows (see Fig. 3.5):

$$d_{3D} = d_{3D-out} + d_{3D-in} = \sqrt{(d_{2D-out} + d_{2D-in})^2 + (h_{gNB} - h_{UE})^2}$$

Fig. 3.6 shows the path loss in dB for the 3D distance from the smallest value supported in each scenario to 10^3 m for the outdoor scenarios and 10^2 m for the indoor cases. In this figure, O2I denotes outdoor-to-indoor loss in various test environments.

3.1.2 Two- and Three-Dimensional Channel Models

The 5G cellular systems are expected to operate over a wide range of frequencies from 450 MHz to 100 GHz. For the development and standardization of the new 5G systems operating in frequency bands above 6 GHz, there was a serious need to accurately model radio signal propagation in these frequency bands, which could not be fully characterized by the existing channel models, because the previous generations of channel models were designed and evaluated for sub-6 GHz frequencies. The development of 3GPP 3D channel

model was a step toward modeling 2D antenna arrays that are used in 5G network deployments. The measurements indicate that the smaller wavelengths increase the sensitivity of the propagation models to the scale of the environment effects and show some frequency dependence of the path loss as well as increased occurrence of blockage phenomenon. Furthermore, the penetration loss is highly dependent on the material and tends to increase with frequency. The shadow fading and angular spread parameters are larger and the boundary between LoS and NLoS depends not only on antenna heights but also on the local environment. The small-scale characteristics of the channel such as delay-spread and angular-spread and the multipath richness is somewhat similar over the, which was a good reason for extending the existing 3GPP models to the wider frequency range [24,25].

The goal of channel modeling is to provide accurate mathematical representations of radio propagation to be used in link-level and system-level simulations corresponding to a specific deployment scenario. Since the radio channel can be assumed as linear, it can be described by its impulse response. Once the impulse response is known, one can determine the response of the radio channel to any input signal. The impulse response is usually represented as a power density function of excess delay, measured relative to the first detectable signal. This function is often referred to as a power delay profile. The channel impulse response varies with the position of the receiver and may also vary with time. Therefore, it is usually measured and reported as an average of profiles measured over one wavelength to reduce noise effects, or over several wavelengths to determine a spatial average. It is important to clarify which average is meant, and how the averaging was performed.

The propagation effects of a wireless channel can be modeled with a large-scale propagation model combined with a small-scale fading model, where the former models long-term slow-fading characteristics of the wireless channel, such as path loss and shadowing, while the small-scale fading model provides rapid fluctuation behavior of the wireless channel due to multipath and Doppler spread. For a wireless channel with multiple antennas, static beamforming gain such as the sectorization beam pattern also contributes to the long-term propagation characteristic of the wireless channel and can be modeled as part of the large-scale propagation model. As for the small-scale fading model of a MIMO channel, the correlation of signals between antenna elements also needs to be considered and can be modeled by a spatial channel model that was used to evaluate performance of the previous generations of cellular standards. Although electromagnetic beam patterns generated by base station antenna arrays are 3D in nature, they were usually modeled as linear horizontal arrays, and elevation angles of signal paths have been ignored for simplicity. The 3D spatial channel models non-zero elevation angles associated with signal paths as well as azimuth angles so that the small-scale fading effect on each antenna element of the 2D antenna grid and the correlation between any two pair of antenna elements on the 2D antenna grid can be modeled.

While the increase in average mutual information (capacity) of a MIMO channel with the number of antennas is well understood, it appears that the variance of the mutual information can grow very slowly or even shrink as the number of antennas increases. This phenomenon is referred to as channel hardening in the literature, which has certain implications for control and data transmission [31].

For a three-sector macrocell, the horizontal and vertical antenna patterns are commonly modeled with a 3 dB beam-width of 70 and 10 degrees, respectively, with an antenna gain of 17 dBi. The vertical antenna pattern is also a function of electrical and mechanical antenna downtilt, where the electrical downtilt is a result of vertical analog beamforming, generated by applying common and static phase shifts to each vertical array of 2D antenna elements. For instance, the electrical downtilt is set to 15 and 6 degrees for a macrocell deployment scenario with inter-site distance of 500 m and 1732 m, respectively, in 3GPP evaluation methodology [4].

In order to meet the technical requirements of IMT-2020, new features are captured in 5G channel models such as support of frequencies up to 100 GHz and large bandwidth, 3D modeling, support of large antenna arrays, blockage modeling, and spatial consistency [4,5]. The 3D modeling describes the channel propagation in azimuth and elevation directions between the transmitting and the receiving antennas. It is more complete and accurate relative to the 2D modeling which only considers the propagation characteristics in the azimuth direction. Multi-antenna techniques capable of exploiting the elevation dimension have been developed for LTE since Rel-13 and are considered very important in 5G, which include modeling the elevation angles of departure and arrival, and their correlation with other parameters [7]. For a base station (access node) equipped with columns of active antenna arrays, vertical analog beamforming can be applied prior to the power amplifier (PA); thus the electrical downtilt can be dynamically adjusted over time. This feature allows the base station to dynamically adjust its cell coverage depending on the user distribution in the cell, for example, analog beamforming can be directed toward UEs congregated at the same location. Nonetheless, such beamforming is still fundamentally cell specific (i.e., spatial separation of individual UE is not possible) and operate on a long timescale, although time-varying base station antennas are modeled having one or multiple antenna panels, where an antenna panel has one or multiple antenna elements placed vertically, horizontally or in a 2D array within each panel. As a result, 3D channel modeling is required for performance evaluation of full-dimension MIMO. The full-dimension MIMO involves precoding/beamforming exploiting both horizontal and vertical degrees of freedom on a small timescale in a frequency-selective manner.

We start our study with the 3D antenna models in mmWave bands where implementation of large antenna arrays is feasible in the gNB or the UE. Let's assume that the gNB and/or

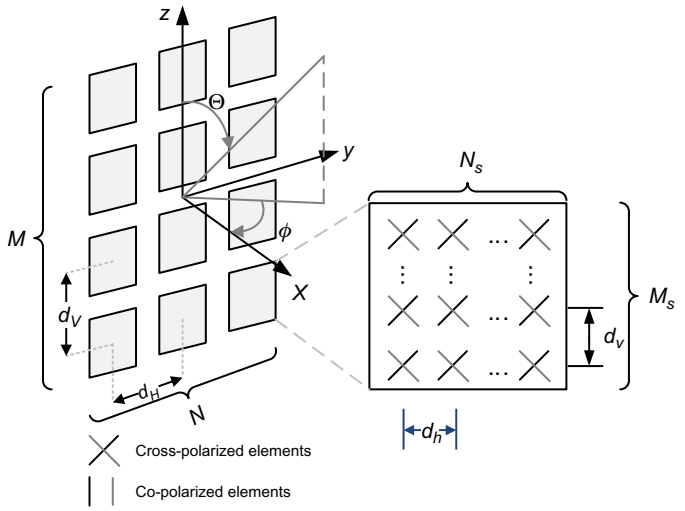


Figure 3.7
3D base station antenna model [4].

the UE each has a 2D planar antenna subarray comprising $M_s \times N_s$ antenna elements, where N_s denotes the number of columns and M_s is the number of antenna elements with the same polarization in each column (see Fig. 3.7). The antenna elements are uniformly spaced with a center-to-center spacing of d_h and d_v in the horizontal and vertical directions, respectively. The $M_s \times N_s$ elements may either be single polarized or dual polarized. A uniform rectangular array is formed comprising $M \times N$ antenna panels where M is the number of panels in a column and N is the number of panels in a row. Antenna panels are uniformly spaced with a center-to-center spacing of d_H and d_V in the horizontal and vertical directions, respectively. The 3GPP 3D channel model allows modeling 2D planar antenna arrays. The antenna elements can either be linearly polarized or cross polarized, as shown in Fig. 3.7. In this regard, the model represents a compromise between practicality and precision as it does not include the mutual coupling effect as well as different propagation effects of horizontally and vertically polarized waves [7].

For each antenna element, the general form of antenna element horizontal radiation pattern can be specified as $A_{E,H}(\phi) = -\min\left(12\left(\phi/\phi_{3\text{dB}}\right)^2, \alpha_m\right)$ where $-180^\text{o} \le \phi \le 180^\text{o}$, $\phi_{3\text{dB}}$ denotes the horizontal 3 dB beam-width, and the parameter α_m is the maximum side-lobe level attenuation.

The general form of antenna element vertical radiation pattern is specified as $A_{E,V}(\theta) = -\min\left(12\left[(\theta-\theta_{tilt})/\theta_{3\text{dB}}\right]^2, \alpha_m\right)$ where $-180^\text{o} \le \theta \le 180^\text{o}$, $\theta_{3\text{dB}}$ denotes the vertical 3 dB beam-width, and θ_{tilt} is the tilt angle. It must be noted that $\theta = 0$ points to the zenith and $\theta = 90^\text{o}$ points to the horizon. The combined vertical and horizontal antenna

element pattern is then given as $A(\theta,\phi) = -\min\left(-\left\lfloor A_{E,V}(\theta) + A_{E,H}(\phi)\right\rfloor, \alpha_m\right)$ where $A(\theta,\phi)$ is the relative antenna gain (dB) of an antenna element in the direction (θ,ϕ) [4,7,24].

The above concepts can also be applied to the UE antenna arrays. In this case $M \times N$ antenna panels may have different orientations. Let $\left(\Omega_{m_g,n_g}, \Theta_{m_g,n_g}\right)$ denote the orientation angles of the panel $\left(m_g, n_g\right)$ $0 \le m_g < M_g, 0 \le n_g < N_g$, where the orientation of the first panel $\left(\Omega_{0,0}, \Theta_{0,0}\right)$ is defined as the UE orientation, Ω_{m_g,n_g} is the array bearing angle, and Θ_{m_g,n_g} is the array downtilt angle. The antenna bearing is defined as the angle between the main antenna lobe center and a line directed toward east. The bearing angle increases in a clockwise direction. The parameters of the base station antenna array pattern for Dense-urban-eMBB (macro-TRP), Rural-eMBB, Urban-macro-mMTC, and Urban-macro-URLLC test environments (i.e., deployment scenarios studied in 3GPP and ITU-R[5]) are defined as $\theta_{3\text{dB}} = \phi_{3\text{dB}} = 65^{\text{o}}$, $\theta_{tilt} = 90^{\text{o}}$, $\alpha_m = 30$ dB. The parameters of the UE antenna pattern for frequencies above 30 GHz are given as $\theta_{3\text{dB}} = \phi_{3\text{dB}} = 90^{\circ}$, $\theta_{tilt} = 90^{\circ}$, *and* $\alpha_m = 25$ dB.

The IMT-2020 3D channel model developed in ITU-R is a geometry-based stochastic channel model. It does not explicitly specify the location of the scatterers, rather the directions of the rays. Geometry-based modeling of the radio channels enables separation of propagation parameters and antennas. The channel parameters for individual snapshots are determined stochastically based on statistical distributions obtained from channel measurements. Channel realizations are generated through the application of the geometrical principle by summing contributions of rays with specific small-scale parameters such as delay, power, azimuth angles of arrival and departure, and elevation angles of arrival and departure. The results are further combined with transmit/receive antenna correlations and temporal fading with geometry-dependent Doppler spectrum effects [4]. A single-link channel model is shown in Fig. 3.8 for the downlink direction. Each circle with several dots represents scattering region creating one cluster, each cluster is constituted by M rays, where a total of N clusters are assumed. We assume that there are N_{tx} antennas at the transmitter and N_{rx} antennas at the receiver. The small-scale parameters such as delay $\tau_{n,m}$, azimuth AoA

[5] Test environment reflects geographic environment and usage scenario which is used for the evaluation process; however, it has a direct relevance to the deployment scenario. The test environments defined in 3GPP and ITU-R are as follows:

- Indoor hotspot-eMBB is an indoor isolated environment at offices and/or in shopping malls based on stationary and pedestrian users with very high user density.
- Dense-urban-eMBB is an urban environment with high user density and traffic loads focusing on pedestrian and vehicular users.
- Rural-eMBB is a rural environment with larger and continuous wide area coverage, supporting pedestrian, vehicular, and high speed vehicular users.
- Urban macro-mMTC is an urban macro-environment targeting continuous coverage focusing on a high number of connected machine type devices.
- Urban macro-URLLC is an urban macro-environment targeting ultra-reliable and low-latency communications.

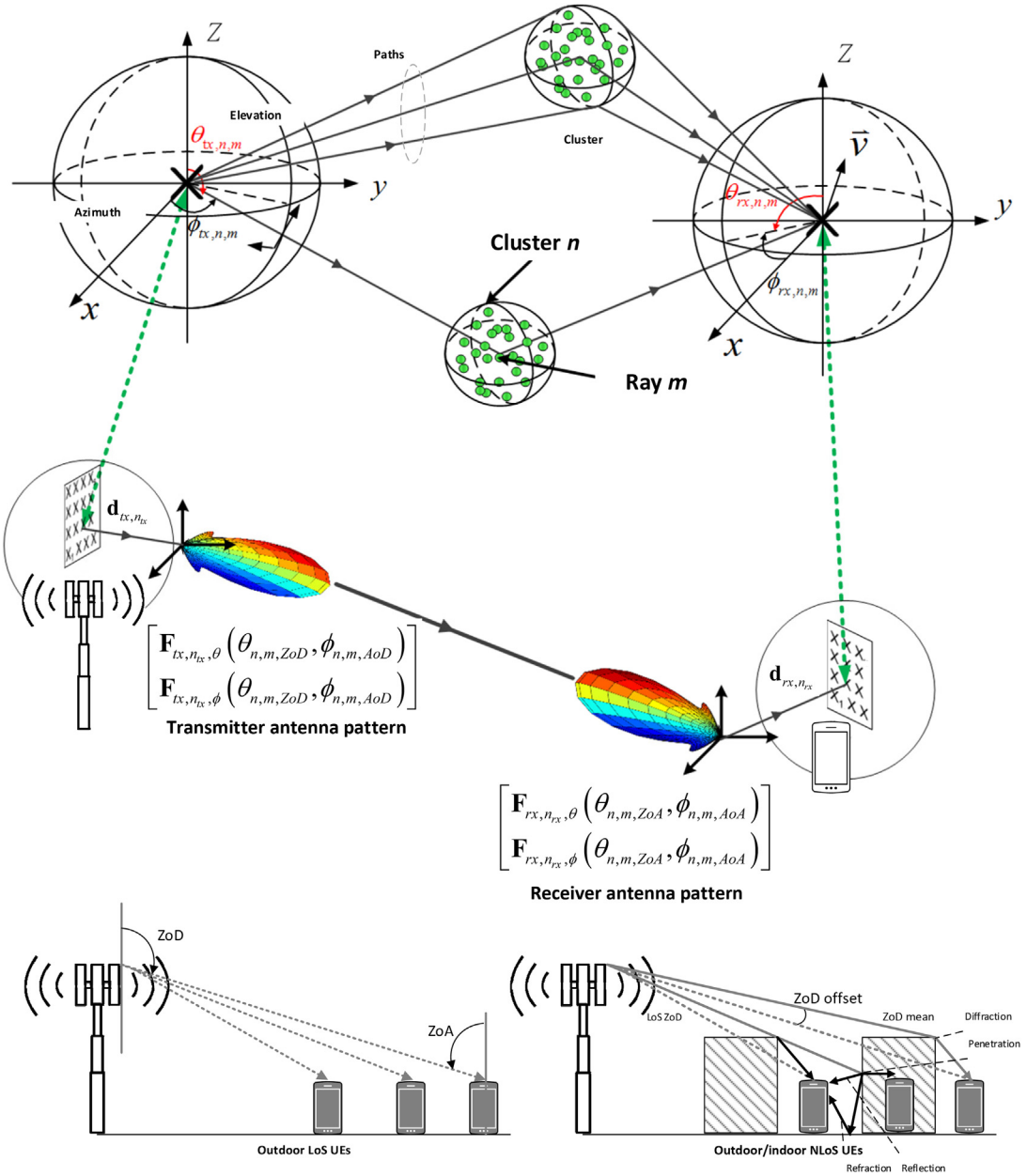


Figure 3.8
Illustration of 3D MIMO channel model and its parameters [4].

$\phi_{rx,n,m}$, elevation AoA $\theta_{rx,n,m}$, azimuth AoD $\phi_{tx,n,m}$ and elevation AoD $\theta_{tx,n,m}$ are assumed to be different for each ray. In the primary module of 3GPP 3D channel model, the number of clusters are fixed and frequency independent. The typical number of clusters reported in the literature is often small, random, and can be modeled as a Poisson distribution. By choosing an appropriate mean value of the Poisson distribution, the events with a larger number of clusters with a low probability may also be produced.

The 3GPP 3D channel model is a geometric stochastic model, describing the scattering environment between base station sector and the UE in both azimuth and elevation dimensions. The scatterers are represented by statistical parameters without having a real physical location. It specifies three propagation conditions, LoS, non-LoS, and outdoor-to-indoor. In each scenario it defines different parameters for mean propagation path loss, macroscopic fading, and microscopic fading. The probability of being in LoS is determined separately for indoor and outdoor UEs depending on the height of the UE as well as the break point distance. The break point distance characterizes the gap between transmitter and receiver at which the Fresnel zone is barely broken for the first time. For an indoor UE, LoS refers to the signal propagation outside of the building in which the UE is located. For each UE location, large-scale parameters are generated according to its geographic position as well as the propagation conditions at this location. The large-scale parameters incorporate shadow fading, the Rician K-factor (only in the LoS case), delay spread, azimuth angle spread of departure and arrival, as well as azimuth angle-spread of departure and arrival. The time-variant impulse response matrix of $N_{rx} \times N_{tx}$ MIMO channel is given by the following expression [4]:

$$\mathbf{H}(t;\tau) = \sum_{n=1}^{N} \mathbf{H}_n(t;\tau)$$

where $t, \tau, N,$ and n denote time, delay, number of clusters, and cluster index, respectively. The above channel impulse response is composed of the antenna array impulse response matrices $\mathbf{F}_{tx}$ and $\mathbf{F}_{rx}$ for the transmitter and the receiver sides, respectively, as well as the dual-polarized propagation channel response matrix. The channel from the transmitter antenna element n_{tx} to the receiver antenna element n_{rx} for the cluster n is given by the following expression [4]:

$$\begin{aligned}
&\mathbf{H}_{n_{rx},n_{tx},n,m}(t;\tau) \\
&= \sqrt{\frac{P_n}{M}} \sum_{m=1}^{M} \begin{bmatrix} \mathbf{F}_{rx,n_{rx},\theta}\left(\theta_{n,m,ZoA}, \phi_{n,m,AoA}\right) \\ \mathbf{F}_{rx,n_{rx},\phi}\left(\theta_{n,m,ZoA}, \phi_{n,m,AoA}\right) \end{bmatrix}^T \begin{bmatrix} \exp\left(j\Phi_{n,m}^{\theta\theta}\right) & \sqrt{\kappa_{n,m}{}^{-1}}\exp\left(j\Phi_{n,m}^{\theta\phi}\right) \\ \sqrt{\kappa_{n,m}{}^{-1}}\exp\left(j\Phi_{n,m}^{\phi\theta}\right) & \exp\left(j\Phi_{n,m}^{\phi\phi}\right) \end{bmatrix} \\
&\quad \times \begin{bmatrix} \mathbf{F}_{tx,n_{tx},\theta}\left(\theta_{n,m,ZoD}, \phi_{n,m,AoD}\right) \\ \mathbf{F}_{tx,n_{tx},\phi}\left(\theta_{n,m,ZoD}, \phi_{n,m,AoD}\right) \end{bmatrix} \exp\left(j2\pi\lambda_0^{-1}\left(\mathbf{r}_{rx,n,m}^T.\mathbf{d}_{rx,n_{rx}}\right)\right) \exp\left(j2\pi\lambda_0^{-1}\left(\mathbf{r}_{tx,n,m}^T.\mathbf{d}_{tx,n_{tx}}\right)\right) \\
&\quad \times \exp\left(j2\pi v_{n,m}t\right)\delta(\tau - \tau_{n,m})
\end{aligned}$$

where P_n is the power of nth path and the other parameters are defined as follows [4]:

- $\Phi_{n,m}^{\theta\theta}$, $\Phi_{n,m}^{\theta\phi}$, $\Phi_{n,m}^{\phi\theta}$, and $\Phi_{n,m}^{\phi\phi}$ denote random initial phase for each ray m of each cluster n and for four different polarization combinations.
- $\mathbf{F}_{rx,n_{rx},\theta}$ and $\mathbf{F}_{rx,n_{rx},\phi}$ represent receive antenna element n_{rx} field patterns in direction of the spherical basis vectors Θ and Φ.
- $\mathbf{F}_{tx,n_{tx},\theta}$ and $\mathbf{F}_{tx,n_{tx},\phi}$ denote transmit antenna element n_{tx} field patterns in direction of the spherical basis vectors Θ and Φ.
- $\mathbf{r}_{rx,n,m}$ and $\mathbf{r}_{tx,n,m}$ denote spherical unit vector with azimuth arrival angle $\phi_{n,m,AoA}$, elevation arrival angle $\theta_{n,m,ZoA}$ and azimuth departure angle $\phi_{n,m,AoD}$, and elevation departure angle $\theta_{n,m,ZoD}$.
- $\mathbf{d}_{rx,n_{rx}}$ and $\mathbf{d}_{tx,n_{tx}}$ represent location vector of receive antenna element n_{rx} and transmit antenna element n_{tx}.
- $\kappa_{n,m}$ is the cross-polarization power ratio in linear scale.
- λ_0 denotes the wavelength of the carrier frequency.
- $v_{n,m}$ represents the Doppler frequency component of ray n, m.

If the radio channel is dynamically modeled, the above small-scale parameters would become time variant. The channel impulse response describes the channel from a transmit antenna element to a receive antenna element. The spherical unit vectors are defined as follows [24]:

$$\hat{\mathbf{r}}_{rx,n,m} = \begin{bmatrix} \sin\theta_{n,m,ZoA} & \cos\varphi_{n,m,AoA} \\ \sin\theta_{n,m,ZoA} & \sin\varphi_{n,m,AoA} \\ \cos\theta_{n,m,ZoA} \end{bmatrix}, \quad \hat{\mathbf{r}}_{tx,n,m} = \begin{bmatrix} \sin\theta_{n,m,ZoD} & \cos\varphi_{n,m,AoD} \\ \sin\theta_{n,m,ZoD} & \sin\varphi_{n,m,AoD} \\ \cos\theta_{n,m,ZoD} \end{bmatrix}$$

The Doppler frequency component $v_{n,m}$ depends on the arrival angles (AoA, ZoA), and the UE velocity vector $\overline{\mathbf{v}}$ with speed v, travel azimuth angle ϕ_v, elevation angle θ_v and is given by

$$v_{n,m} = \frac{\hat{\mathbf{r}}_{rx,n,m}^{T} \cdot \overline{\mathbf{v}}}{\lambda_0}, \quad \overline{\mathbf{v}} = v\left[\sin\theta_v\cos\varphi_v \;\; \sin\theta_v\sin\varphi_v \;\; \cos\theta_v\right]^T$$

The parameters $\mathbf{d}_{rx,n_{rx}}$ and $\mathbf{d}_{tx,n_{tx}}$ are the location vectors of receive and transmit antenna elements, respectively. Considering a base station sector with coordinates (s_x, s_y, s_z) and a planar antenna array, the location vector per antenna element is defined as $\mathbf{d}_{tx,n_{tx}} = \begin{pmatrix} s_x & s_y & s_z \end{pmatrix}^T + \begin{pmatrix} 0 & (k-1)d_H & (l-1)d_V \end{pmatrix}^T$ where $k = 1, 2, \ldots, N$ and $l = 1, 2, \ldots, M$ (see Fig. 3.7). The parameters N and d_H denote the number of antenna elements and the inter-element spacing in the horizontal direction, respectively, while M and d_V represent the number of antenna elements and the inter-element spacing in the vertical direction, respectively. The Doppler frequency component of the UE moving at velocity $\mathbf{v}$ is represented by parameter $v_{n,m}$.

When antenna arrays are deployed at the transmitter and receiver, the impulse response of such arrangement results in a vector channel. An example of this configuration is given below for the case of a 2D antenna array [4]:

$$\mathbf{H}_n^{NLoS}(t) = \sqrt{\frac{P_n}{M}} \sum_{m=1}^{M} \begin{bmatrix} F_{rx,\theta}\left(\theta_{n,m,ZoA}, \phi_{n,m,AoA}\right) \\ F_{rx,\phi}\left(\theta_{n,m,ZoA}, \phi_{n,m,AoA}\right) \end{bmatrix}^T \begin{bmatrix} \exp\left(j\Phi_{n,m}^{\theta\theta}\right) & \sqrt{\kappa_{n,m}^{-1}}\exp\left(j\Phi_{n,m}^{\theta\phi}\right) \\ \sqrt{\kappa_{n,m}^{-1}}\exp\left(j\Phi_{n,m}^{\phi\theta}\right) & \exp\left(j\Phi_{n,m}^{\phi\phi}\right) \end{bmatrix}$$
$$\times \begin{bmatrix} F_{tx,\theta}\left(\theta_{n,m,ZoD}, \phi_{n,m,AoD}\right) \\ F_{tx,\phi}\left(\theta_{n,m,ZoD}, \phi_{n,m,AoD}\right) \end{bmatrix} \mathbf{a}_{rx}\left(\theta_{n,m,ZoA}, \phi_{n,m,AoA}\right) \mathbf{a}_{tx}^H\left(\theta_{n,m,ZoD}, \phi_{n,m,AoD}\right) \exp\left(j2\pi\nu_{n,m}t\right)$$

$$\mathbf{H}_n^{LoS}(t) = \begin{bmatrix} F_{rx,\theta}\left(\theta_{LoS,ZoA}, \phi_{LoS,AoA}\right) \\ F_{rx,\phi}\left(\theta_{LoS,ZoA}, \phi_{LoS,AoA}\right) \end{bmatrix}^T \begin{bmatrix} \exp(j\Phi_{LoS}) & 0 \\ 0 & \exp(j\Phi_{LoS}) \end{bmatrix}$$
$$\times \begin{bmatrix} F_{tx,\theta}\left(\theta_{LoS,ZoD}, \phi_{LoS,AoD}\right) \\ F_{tx,\phi}\left(\theta_{LoS,ZoD}, \phi_{LoS,AoD}\right) \end{bmatrix} \mathbf{a}_{rx}\left(\theta_{LoS,ZoA}, \phi_{LoS,AoA}\right) \mathbf{a}_{tx}^H\left(\theta_{LoS,ZoD}, \phi_{LoS,AoD}\right) \exp(j2\pi\nu_{LoS}t)$$

where $\mathbf{a}_{tx}\left(\theta_{n,m,ZoD}, \phi_{n,m,AoD}\right)$ and $\mathbf{a}_{rx}\left(\theta_{n,m,ZoA}, \phi_{n,m,AoA}\right)$ are the transmit and receive antenna array impulse response vectors, respectively, corresponding to ray $m \in \{1, \ldots, M\}$ in cluster $n \in \{1, \ldots, N\}$ given by the following expression:

$$\mathbf{a}_{tx}\left(\theta_{n,m,ZoD}, \phi_{n,m,AoD}\right) = \exp\left(j\frac{2\pi}{\lambda_0}\left[\mathbf{W}_{tx}\mathbf{r}_{tx}\left(\theta_{n,m,ZoD}, \phi_{n,m,AoD}\right)\right]\right); \quad \forall n, m$$
$$\mathbf{a}_{rx}\left(\theta_{n,m,ZoA}, \phi_{n,m,AoA}\right) = \exp\left(j\frac{2\pi}{\lambda_0}\left[\mathbf{W}_{rx}\mathbf{r}_{rx}\left(\theta_{n,m,ZoA}, \phi_{n,m,AoA}\right)\theta\right]\right); \quad \forall n, m$$

where λ_0 is the wavelength of carrier frequency f_0; $\mathbf{r}_{tx}(\theta_{n,m,ZoD}, \phi_{n,m,AoD})$ and $\mathbf{r}_{rx}(\theta_{n,m,ZoA}, \phi_{n,m,AoA})$ are the corresponding angular 3D spherical unit vectors of the transit and receive, respectively; $\mathbf{W}_{tx}$ and $\mathbf{W}_{rx}$ denote the location matrices of the transmit and receive antenna elements in 3D Cartesian coordinates. The location matrices in the vectored impulse response are provided for an antenna configuration that is a uniform rectangular array consisting of cross-polarized antenna elements shown in Fig. 3.4.

The Doppler shift generally depends on the time-variance of the channel as it is defined as the derivative of the channel phase over time. It can result from transmitter, receiver, or scatterers movement. The general form of the exponential Doppler component is given as follows [24]:

$$\nu_{n,m} = \exp\left(j2\pi\int_{t_0}^{t}\frac{\hat{\mathbf{r}}_{rx,n,m}^T(\tilde{t})\cdot\mathbf{v}(\tilde{t})}{\lambda_0}d\tilde{t}\right)$$

where $\hat{\mathbf{r}}_{rx,n,m}(t)$ denotes the normalized vector that points to the direction of the incoming wave as seen from the receiver at time t. The velocity vector of the receiver at time t is

denoted by $\mathbf{v}(t)$ while t_0 denotes a reference point in time that defines the initial phase $t_0 = 0$. The above expression is only valid for time-invariant Doppler shift, satisfying $\hat{\mathbf{r}}_{rx,n,m}^{T}(t) \cdot \mathbf{v}(t) = \hat{\mathbf{r}}_{rx,n,m}^{T} \cdot \mathbf{v}$.

Spatial correlation is often said to degrade the performance of multi-antenna systems and imposes a limit on the number of antennas that can be effectively fit in a small mobile device. This seems intuitive as the spatial correlation decreases the number of independent channels that can be created by precoding. When modeling spatial correlation, it is useful to employ the Kronecker model, where the correlation between transmit and receive antennas are assumed independent and separable. This model is reasonable when the main scattering appears close to the antenna arrays and has been validated by outdoor and indoor measurements. Assuming Rayleigh fading, the Kronecker model means that the channel matrix can be represented as $\mathbf{H} = \mathbf{R}_{RX}^{1/2}\mathbf{H}_{channel}\left(\mathbf{R}_{TX}^{1/2}\right)^{T}$ where the elements of $\mathbf{H}_{channel}$ are independent and identically distributed as circular symmetric complex Gaussian random variables with zero-mean and unit variance. The important part of the model is that $\mathbf{H}_{channel}$ is pre-multiplied by the receive-side spatial correlation matrix $\mathbf{R}_{RX}$ and post-multiplied by transmit-side spatial correlation matrix $\mathbf{R}_{TX}$. Equivalently, the channel matrix can be expressed as $\mathbf{H} \sim \mathbb{C}(0, \mathbf{R}_{TX} \otimes \mathbf{R}_{RX})$ where $\otimes$ denotes the Kronecker product [9].

Using the Kronecker model, the spatial correlation depends directly on the eigenvalue distributions of the correlation matrices $\mathbf{R}_{RX}$ and $\mathbf{R}_{TX}$. Each eigenvector represents a spatial direction of the channel and its corresponding eigenvalue describes the average channel/signal gain in that direction. For the transmit-side, the correlation matrix $\mathbf{R}_{TX}$ describes the average gain in a spatial transmit direction, while receive-side correlation matrix $\mathbf{R}_{RX}$ describes a spatial receive direction. High spatial correlation is represented by large eigenvalue spread in $\mathbf{R}_{TX}$ or $\mathbf{R}_{RX}$, implying that some spatial directions are statistically stronger than the others. On the other hand, low spatial correlation is represented by small eigenvalue spread in $\mathbf{R}_{TX}$ or $\mathbf{R}_{RX}$, which implies that almost the same signal gain can be expected from all spatial directions [9].

Let's now focus on a specific case of downlink transmission from a base station to a mobile station. Denoting the nth snapshot of the spatial correlation matrices at the gNB and the UE by $\mathbf{R}_{gNB,n}$ and $\mathbf{R}_{UE,n}$, the per-tap spatial correlation is determined as the Kronecker product[6] of the gNB and UE's antenna correlation matrices as $\mathbf{R}_n = \mathbf{R}_{gNB,n} \otimes \mathbf{R}_{UE,n}$. We denote the number of receive antennas by N_{RX} and the number of transmit antennas by N_{TX}. If

[6] Given a $m \times n$ matrix $\mathbf{A}$ and a $p \times q$ matrix $\mathbf{B}$, the Kronecker product $\mathbf{C} = \mathbf{A} \otimes \mathbf{B}$, also called matrix direct product, is an $mp \times nq$ matrix with elements defined by $c_{\alpha\beta} = a_{ij}b_{kl}$ where $\alpha \equiv p(i-1)+k$ and $\beta \equiv q(j-1)+l$. The matrix direct product provides the matrix of the linear transformation induced by the vector space tensor product of the original vector spaces. More precisely, suppose that operators $S{:}V_1 \to W_1$ and $T{:}V_2 \to W_2$ are given by $S(x) = Ax$ and $T(y) = By$, then $S \otimes T{:}V_1 \otimes V_2 \to W_1 \otimes W_2$ is determined by $S \otimes T(x \otimes y) = (Ax) \otimes (By) = (A \otimes B)(x \otimes y)$.

Table 3.3: gNB/UE correlation matrix [9].

Entity	Number of Antennas		
	1	2	4
gNB	$\mathbf{R}_{gNB} = 1$	$\mathbf{R}_{gNB} = \begin{pmatrix} 1 & \alpha \\ \alpha^* & 1 \end{pmatrix}$	$\mathbf{R}_{gNB} = \begin{pmatrix} 1 & \alpha^{1/9} & \alpha^{4/9} & \alpha \\ \alpha^{1/9*} & 1 & \alpha^{1/9} & \alpha^{4/9} \\ \alpha^{4/9*} & \alpha^{1/9*} & 1 & \alpha^{1/9} \\ \alpha^* & \alpha^{4/9*} & \alpha^{1/9*} & 1 \end{pmatrix}$
UE	$\mathbf{R}_{UE} = 1$	$\mathbf{R}_{UE} = \begin{pmatrix} 1 & \beta \\ \beta^* & 1 \end{pmatrix}$	$\mathbf{R}_{UE} = \begin{pmatrix} 1 & \beta^{1/9} & \beta^{4/9} & \beta \\ \beta^{1/9*} & 1 & \beta^{1/9} & \beta^{4/9} \\ \beta^{4/9*} & \beta^{1/9*} & 1 & \beta^{1/9} \\ \beta^* & \beta^{4/9*} & \beta^{1/9*} & 1 \end{pmatrix}$

Table 3.4: Values of α and β for different antenna correlations [9].

Low Correlation		Medium Correlation		High Correlation	
α	β	α	β	α	β
0	0	0.9	0.3	0.9	0.9

cross-polarized antennas are present at the receiver, it is assumed that $N_{RX}/2$ receive antennas have the same polarization, while the remaining $N_{RX}/2$ receive antennas have orthogonal polarization. Likewise, if cross-polarized antennas are present at the transmitter, it is assumed that $N_{TX}/2$ transmit antennas have the same polarization, while the remaining $N_{TX}/2$ transmit antennas have orthogonal polarization. It is further assumed that the antenna arrays are composed of pairs of co-located antennas with orthogonal polarization. Under these assumptions, the per-tap channel correlation is determined as $\mathbf{R}_n = \mathbf{R}_{gNB,n} \otimes \mathbf{\Gamma} \otimes \mathbf{R}_{UE,n}$ where $\mathbf{R}_{UE,n}$ is an $N_{RX} \times N_{RX}$ matrix, if all receive antennas have the same polarization, or an $N_{RX}/2 \times N_{RX}/2$ matrix, if the receive antennas are cross-polarized. Likewise, $\mathbf{R}_{gNB,n}$ is an $N_{TX} \times N_{TX}$ matrix, if all transmit antennas have the same polarization, or a $N_{TX}/2 \times N_{TX}/2$ matrix, if the transmit antennas are cross-polarized. Matrix $\mathbf{\Gamma}$ is a cross-polarization matrix based on the cross-polarization defined in the cluster-delay-line models. Matrix $\mathbf{\Gamma}$ is a 2×2 matrix, if cross-polarized antennas are used at the transmitter or at the receiver. It is a 4×4 matrix if cross-polarized antennas are used at both the transmitter and the receiver [9]. Table 3.3 defines the correlation matrices for the gNB and UE in NR with different number of transmit/receive antennas at each entity [9].

For the scenarios with more antennas at either gNB, UE, or both, the channel spatial correlation matrix can be expressed as the Kronecker product of $\mathbf{R}_{gNB}$ and $\mathbf{R}_{UE}$ according to $\mathbf{R}_{spatial} = \mathbf{R}_{UE} \otimes \mathbf{R}_{gNB}$. The parameters α and β for different antenna correlation types are given in Table 3.4. The 3D channel model parameters and model generation procedure are summarized in Fig. 3.9.

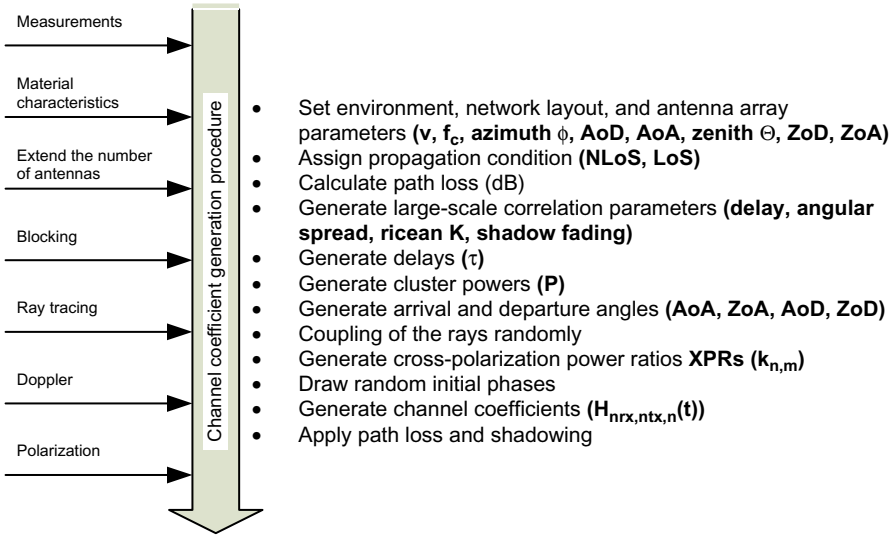


Figure 3.9
Summary of 3D channel model parameters and model generation procedure [7].

3.2 Waveforms

There have been considerable discussions on whether a new type of transmission waveform besides the incumbent cyclic prefix (CP)-OFDM should be adopted for NR [10,11]. Several alternative OFDM-based waveforms, including filter bank multicarrier and generalized frequency division multiplexing, were studied. Many of them claimed advantages in terms of increased bandwidth efficiency, relaxed synchronization requirements, reduced inter-user interference, reduced out-of-band (OOB) emissions, and so on, but at the same time created challenges in terms of increased transceiver complexity, difficulty in MIMO integration, and significant specification impacts. This section describes some of the prominent 5G waveform candidates and their characteristics as well as the reasons that the status-quo OFDM waveform continued to be supported in the new radio. Furthermore, the waveform, numerology, and frame structure should be chosen to enable efficient time/frequency utilization for frequency division duplex (FDD) and time division duplex (TDD) deployments, respectively. Table 3.5 summarizes the design requirements concerning the choice of the waveform, which were used to examine the waveform candidates for the NR.

Gabor's theory of communication suggests that ideally a multicarrier system such as OFDM must satisfy the following requirements [45]:

- The subcarriers are mutually orthogonal in time and frequency to make the receiver as simple as possible and to maintain the inter-carrier interference as low as possible.

Table 3.5: Summary of design targets for the waveform [45,47].

Design Criteria	Remarks
Higher spectral efficiency and scalability	High spectral efficiency for high data rates and efficient use of the available spectrum Ability to efficiently support MIMO and multipath robustness Low latency
Lower in-band and out-of-band emissions	Reduce interference among users within allocated band and reduce interference among neighbor operators
Enables asynchronous multiple access	Support a higher number of small-cell data burst devices with minimal scheduling overhead through asynchronous operations and enables lower power operation
Lower power consumption	Low peak-to-average power ratio allowing efficient power amplifier design
Lower implementation complexity	Reasonable transmitter and receiver complexity and additional complexity must be justified by significant performance improvements
Coexistence with legacy and mobility Support	Simplify LTE coexistence Robust against Doppler shift to allow high mobility

- The transmission waveform is well localized in time and frequency. This provides immunity to ISI from multipath propagation or delay spread and to ICI from Doppler spread. A good time localization is required to enable low latency.
- Maximal spectral efficiency, that is, $\rho = (\delta_T \delta_F)^{-1}$ with ρ denoting the spectral efficiency in data symbols per second per Hertz.

However, it is shown in the literature [38,46], that it is not possible to satisfy these three requirements at the same time and certain tradeoffs are necessary. This conclusion has an impact on the waveform selection in wireless communication systems.

3.2.1 OFDM Basics and Transmission Characteristics

OFDM was selected in LTE/LTE-A due to its efficiency and simplicity using baseband modulation and demodulation stages based on FFT. Mathematically, the nth OFDM symbol can be described as

$$x_n(t) = \sum_{k=0}^{N_{FFT}-1} s_{k,n}\text{rect}(t - nT_u)e^{j(2\pi/T_u)kt}$$

where k denotes the subcarrier index and n denotes symbol index, N_{FFT} is the total number of subcarriers, rect(.) is a rectangular pulse with the symbol period of T_u, and $s_{k,n}$ is the data symbol [e.g., quadrature amplitude modulation (QAM) symbol] of the kth subcarrier at the nth time instant. A CP is appended to the beginning of each symbol. The inserted CP serves as a guard time between symbols which protects against ISI. In addition, it preserves the orthogonality between subcarriers after passing through a channel provided that the CP

duration is longer than the channel RMS delay spread. The CP acts as a buffer region where delayed information from the previous symbols can be stored. The receiver must exclude samples from the CP which might be corrupted by the previous symbol when choosing the samples for an OFDM symbol. When demodulating the received symbol, the receiver can choose T_u/T_s samples from a region which is not affected by the previous symbol.

In a conventional serial data transmission system the information bearing symbols are transmitted sequentially, with the frequency spectrum of each symbol occupying the entire available bandwidth. An unfiltered QAM signal spectrum can be described in the form of $\sin(\pi f T_u)/\pi f T_u$ with zero-crossing points at integer multiples of $1/T_u$, where T_u is the QAM symbol period. The concept of OFDM is to transmit the data bits in parallel QAM-modulated subcarriers using frequency division multiplexing. The carrier spacing is carefully selected so that each subcarrier is located on other subcarriers' zero-crossing points in the frequency domain. Although there are spectral overlaps among subcarriers, they do not interfere with each other, if they are sampled at the subcarrier frequencies. In other words, they maintain spectral orthogonality. The OFDM signal in frequency domain is generated through aggregation of N_{FFT} parallel QAM-modulated subcarriers where adjacent subcarriers are separated by subcarrier spacing $1/T_u$. Since an OFDM signal consists of many parallel QAM subcarriers, the mathematical expression of the signal in time domain can be expressed as follows:

$$x(t) = Re\left\{ e^{j\omega_c t} \sum_{k=-(N_{FFT}-1)/2}^{(N_{FFT}-1)/2} s_k e^{j2\pi k(t-t_g)/T_u} \right\}, \quad nT_u \le t \le (n+1)T_u$$

where $x(t)$ denotes the OFDM signal in time domain, s_k is the complex-valued data that is QAM-modulated and transmitted over subcarrier k, N_{FFT} is the number of subcarriers in frequency domain, ω_c is the RF carrier frequency, and T_g is the guard interval or the CP length. For a large number of subcarriers, direct generation and demodulation of the OFDM signal would require arrays of coherent sinusoidal generators which can become excessively complex and expensive. However, one can notice that the OFDM signal is actually the real part of the inverse discrete Fourier transform (IDFT) of the original complex-valued data symbols $\{s_k | k = -(N_{FFT}-1)/2, \ldots, (N_{FFT}-1)/2\}$. It can be observed that there are $N < N_{FFT}$ subcarriers each carrying the corresponding data α_k. The inverse of the subcarrier spacing $\Delta f = 1/T_u$ is defined as the OFDM useful symbol duration T_u, which is N_{FFT} times longer than that of the original input data symbol duration.

3.2.1.1 Cyclic Prefix

The inclusion of CP in OFDM makes it robust to timing synchronization errors. Robustness to synchronization errors is relevant when synchronization is hard to achieve such as over

the sidelink. It can also be relevant if asynchronous transmissions are allowed in the uplink. The inclusion of he cyclic prefix adds redundancy to the transmission since the same content is transmitted twice as the CP is a copy of the tail of a symbol placed at its beginning. This overhead can be expressed as a function of symbol duration and duration of the CP as $OH = T_{CP}/(T_{CP} + T_u)$. OFDM is a flexible waveform that can support diverse services in a wide range of frequencies when properly selecting subcarrier spacing and CP. Further discussion on OFDM numerology design that fulfills a wide range of requirements is given in the next section.

Since IDFT is used in the OFDM modulator, the original data are defined in the frequency domain, while the OFDM signal $s(t)$ is defined in the time domain. The IDFT can be implemented via a computationally efficient FFT algorithm. The orthogonality of subcarriers in OFDM can be maintained and individual subcarriers can be completely separated and demodulated by an FFT at the receiver when there is no ISI introduced by communication channel. In practice, linear distortions such as multipath delay cause ISI between OFDM symbols, resulting in loss of orthogonality and an effect that is similar to cochannel interference. However, when delay spread is small, that is, within a fraction of the OFDM useful symbol length, the impact of ISI is negligible, although it depends on the order of modulation implemented by the subcarriers (see Fig. 3.10). A simple solution to mitigate multipath delay is to increase the OFDM effective symbol duration such that it is much larger than the delay spread; however, when the delay spread is large, it requires a large number of subcarriers and a large FFT size. Meanwhile, the system might become sensitive to Doppler shift and carrier frequency offset. An alternative approach to mitigate multipath distortion is to generate a cyclically extended guard interval, where each OFDM symbol is prefixed with a periodic extension of the signal itself, as shown in Fig. 3.10 where the tail of the symbol is copied to the beginning of the symbol. The OFDM symbol duration then is defined as $T_s = T_u + T_g$, where T_g is the guard interval or CP. When the guard interval is longer than the channel impulse response or the multipath delay, the ISI can be effectively eliminated. The ratio of the guard interval to useful OFDM symbol duration depends on the deployment scenario and the frequency band. Since the insertion of the guard intervals will reduce the system throughput, T_g is usually selected less than $T_u/4$. The CP should absorb most of the signal energy dispersed by the multipath channel. The entire the ISI energy is contained within the CP, if its length is greater than that of the channel RMS delay spread $\left(T_g > \tau_{RMS}\right)$. In general it is sufficient to have most of the delay spread energy absorbed by the guard interval, considering the inherent robustness of large OFDM symbols to time dispersion. Fig. 3.11 illustrate the OFDM modulation and demodulation process in the transmitter and the receiver, respectively. In practice, a windowing or filtering scheme is utilized in the OFDM transmitter side to reduce the OOB emissions of the OFDM signal (see Fig. 3.10).

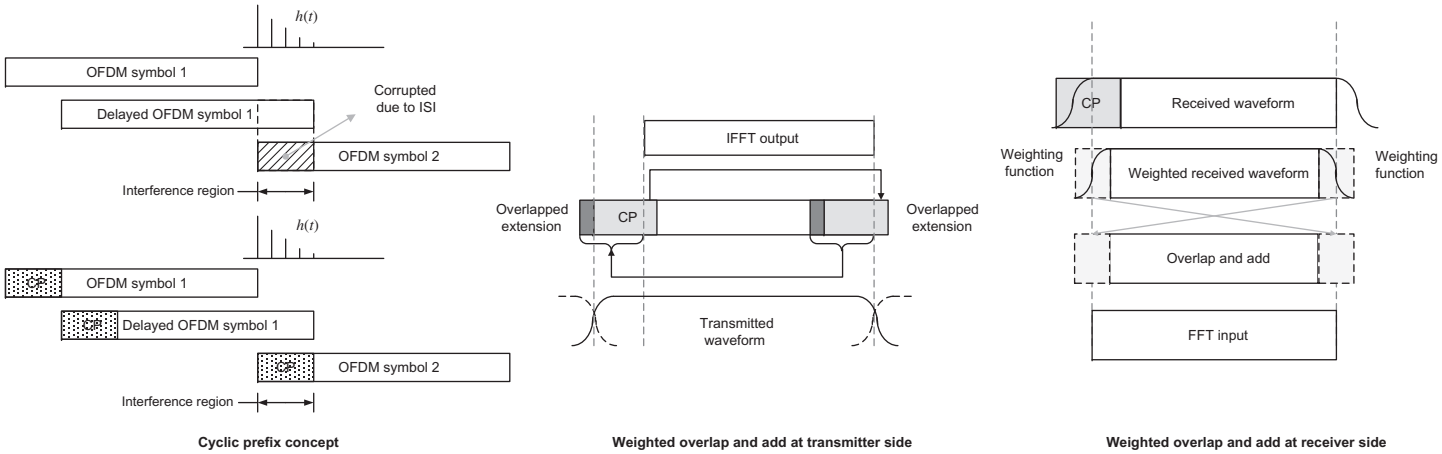


Figure 3.10
Illustration of the effect of cyclic prefix for eliminating ISI ($h(t)$ is the hypothetical channel impulse response) and practical implementation of CP insertion and removal [47].

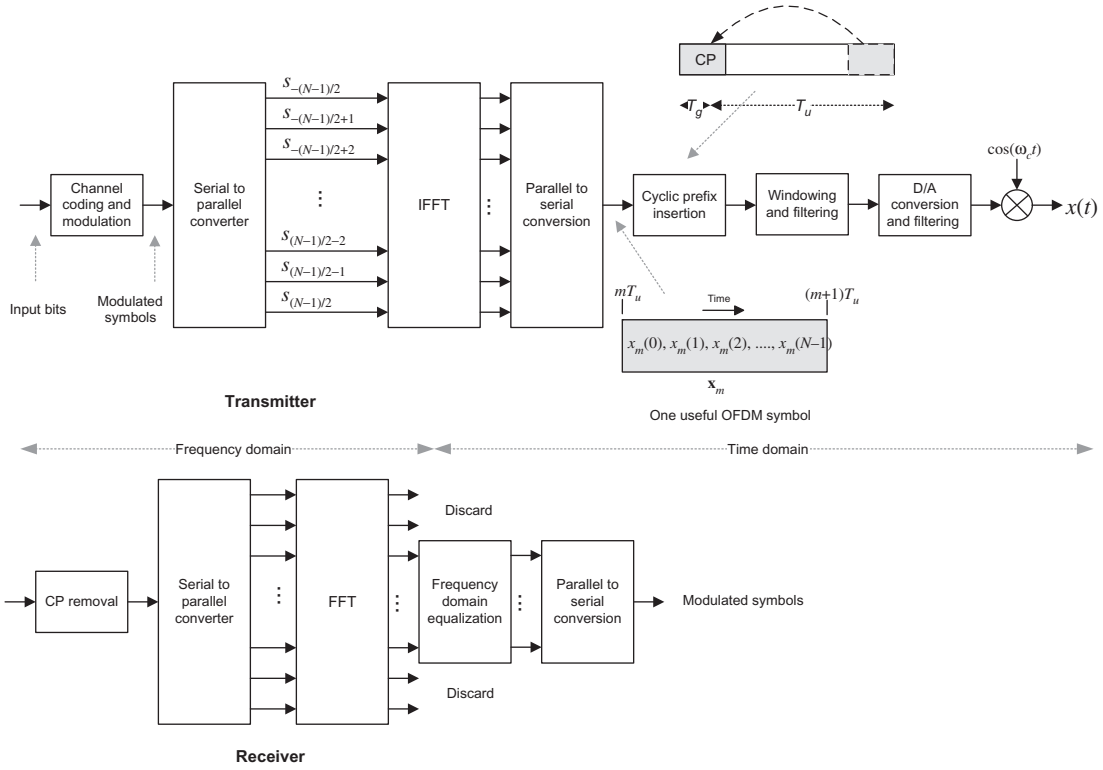


Figure 3.11
OFDM signal generation and reception process.

The mapping of the modulated data symbol into multiple subcarriers also allows an increase in the symbol duration. The symbol duration obtained through an OFDM scheme is much larger than that of a single-carrier modulation technique with a similar transmission bandwidth. In general, when the channel delay spread exceeds the guard time, the energy contained in the ISI will be much smaller with respect to the useful OFDM symbol energy, as long as the symbol duration is much larger than the channel delay spread, that is, $T_s \gg \tau_{RMS}$. Although large OFDM symbol duration is desirable to mitigate the ISI effects caused by time dispersion, large OFDM symbol duration can further reduce the ability to alleviate the effects of fast fading, particularly, if the symbol period is large compared to the channel coherence time, then the channel can no longer be considered as time-invariant over the OFDM symbol duration; therefore, this will introduce the inter-subcarrier orthogonality loss. This can affect the performance in fast fading conditions. Hence, the symbol duration should be kept smaller than the minimum channel coherence time. Since the channel

coherence time is inversely proportional to the maximum Doppler spread, the symbol duration T_s must, in general, be chosen such that $T_s \ll 1/\nu_{RMS}$.

The large number of OFDM subcarriers makes the bandwidth of the individual subcarriers small relative to the overall signal bandwidth. With an adequate number of subcarriers, the inter-carrier spacing is much smaller than the channel coherence bandwidth. Since the channel coherence bandwidth is inversely proportional to the channel delay spread τ_{RMS}, the subcarrier separation is generally designed such that $\Delta f \ll 1/\tau_{RMS}$. In this case, the fading on each subcarrier is flat and can be modeled as a complex-valued constant channel gain. The individual reception of the modulated symbols transmitted on each subcarrier is therefore simplified to the case of a flat-fading channel. This enables a straightforward introduction of advanced MIMO schemes. Furthermore, in order to mitigate Doppler spread effects, the inter-carrier spacing should be much larger than the RMS Doppler spread $\Delta f \gg \nu_{RMS}$. Since the OFDM sampling frequency is typically larger than the actual signal bandwidth, only a subset of subcarriers is used to carry modulated symbols. The remaining subcarriers are left inactive prior to the IFFT and are referred to as guard subcarriers. The split between the active and the inactive subcarriers is determined based on the spectral sharing and regulatory constraints, such as the bandwidth allocation and the spectral mask. An OFDM transmitter diagram is shown in Fig. 3.11. The incoming bit stream is channel coded and modulated to form the complex-valued modulated symbols. The modulated symbols are converted from serial to parallel with $N < N_{FFT}$ complex-valued numbers per block, where N_{FFT} is the size of FFT/IFFT operation. Each block is processed by an IFFT and the output of the IFFT forms an OFDM symbol, which is converted back to serial data for transmission. A guard interval or CP is inserted between symbols to eliminate ISI effects caused by multipath distortion. The discrete symbols are windowed/filtered and converted to an analog signal for RF upconversion. The reverse process is performed at the receiver. A one-tap equalizer is usually used for each subcarrier to correct channel distortion. The tap coefficients are calculated based on channel information.

When there is multipath distortion, a conventional single-carrier wideband transmission system suffers from frequency-selective fading. A complex adaptive equalizer must be used to equalize the in-band fading. The number of taps required for the equalizer is proportional to the symbol rate and the multipath delay. For an OFDM system, if the guard interval is larger than the multipath delay, the ISI can be eliminated and orthogonality can be maintained among subcarriers. Since each OFDM subcarrier occupies a very narrow spectrum, in the order of a few kHz, even under severe multipath distortion, they are only subject to flat fading. In other words, the OFDM converts a wideband frequency-selective fading channel to a series of narrowband frequency non-selective fading subchannels by using the parallel multicarrier transmission scheme. Since OFDM data subcarriers are statistically independent and identically distributed, based on the central-limit theorem, when the number of subcarriers N_{FFT} is large, the OFDM signal distribution tends to be Gaussian.

3.2.1.2 Pre- and Post-processing Signal-to-Noise Ratio

In an OFDM system, the SNR is a measure of channel quality and is a key factor of link-level error assessment. There are different methods for calculation of SNR in single-antenna and multi-antenna transmission systems. For single-input/single-output systems, the SNR can be viewed as the received SNR, that is, the received SNR before the detector. The post-processing SNR is often used for MIMO links and represents the SNR after combining in the receiver and measures the likelihood that a coded message is decoded successfully. In link-level simulations, the SNR γ is typically calculated using the following method. Let's vector $\mathbf{x} = (x_1, x_2, \ldots, x_{N_{TX}})^H \in \mathbb{C}^{N_{TX} \times 1}$ denotes the transmit signal where $x_k \in \mathbb{C} \forall k = 1, 2, \ldots, N_{TX}$ is the complex-valued transmitted symbols from the kth transmit antenna. Note that N_{TX} is the number of transmit antennas. It can be shown that the total transmit signal power can be obtained as $\sigma_{\mathbf{x}}^2 = \text{trace}(\mathbf{R}_{\mathbf{x}})$ where $\mathbf{R}_{\mathbf{x}}$ denotes the autocorrelation matrix of the transmitted signal. The transmit power from the kth antenna is given as $\sigma_{x_k}^2 = E\{|x_k|^2\} = 1/N_{TX}$ (uniformly distributed power). If $\mathbf{H}$ represents the channel matrix with $\|\mathbf{H}\|_F^2 = N_{TX}N_{RX}$[7] in which N_{TX} and N_{RX} denote the number of transmit and receive antennas, respectively, the received signal vector $\mathbf{y}$ can be calculated as $\mathbf{y} = \mathbf{Hx} + \mathbf{v}$ where complex-valued Gaussian-distributed noise vector $\mathbf{v} \sim \mathbb{C}(\mathbf{0}, \sigma_{\mathbf{v}}^2\mathbf{I})$ denotes the noise vector with respect to the size of the FFT N_{FFT} and the number of used subcarriers before the detector N_{used}. We define the complex-valued Gaussian-distributed noise vector $\mathbf{n} \sim \mathbb{C}(\mathbf{0}, \sigma_{\mathbf{n}}^2\mathbf{I})$ to be noise after the FFT operation. The receive SNR before the detector is given as $\gamma_{pre-FFT} = \left(\|\mathbf{Hx}\|_F^2 / N_{RX}\sigma_{\mathbf{v}}^2\right) = \sigma_{\mathbf{v}}^{-2}$, whereas the SNR after the FFT operation is given as $\gamma_{post-FFT} = \left(\|\mathbf{Hx}\|_F^2 / N_{RX}\sigma_{\mathbf{n}}^2\right) = \sigma_{\mathbf{n}}^{-2}$. It can be observed that the difference between pre-FFT and post-FFT SNRs $\left(\gamma_{pre-FFT}/\gamma_{post-FFT}\right) = \left(\sigma_{\mathbf{v}}^2/\sigma_{\mathbf{n}}^2\right) = \left(N_{FFT}/N_{used}\right)$; $N_{used} \leq N_{FFT}$ is always positive, implying that the FFT operation suppresses the noise and enhances the SNR.

3.2.1.3 Peak-to-Average Power Ratio

The peak-to-average power ratio (PAPR) for a single-carrier modulation signal depends on its constellation and the pulse-shaping filter roll-off factor. For a Gaussian distributed

[7] The p-norm of matrix $\mathbf{H}$ is defined as $\|\mathbf{H}\|_p = \left(\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}|h_{ij}|^p\right)^{1/p}$. In the special case when $p = 2$ the norm is called the Frobenius norm and for $p = \infty$ is called the maximum norm. The Frobenius norm or Hilbert–Schmidt norm of matrix $\mathbf{H}$ is similar, although the latter term is often reserved for the operators on a Hilbert space. In general, this norm can be defined in the following forms: $\|\mathbf{H}\|_p = \sqrt{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}|h_{ij}|^2} = \sqrt{\text{trace}(\mathbf{H}^H\mathbf{H})} = \sqrt{\sum_{i=0}^{\min(M,N)-1}\zeta_i^2}$ where $\mathbf{H}^H$ denotes the conjugate transpose of matrix $\mathbf{H}$ and ζ_i are the singular values of matrix $\mathbf{H}$. The Frobenius norm is further similar to the Euclidean norm on $\mathbb{R}^N$ and is obtained from an inner product on the space of all matrices. The Frobenius norm is submultiplicative and is very useful for numerical linear algebra.

OFDM signal, the cumulative distribution function (CDF)[8] of PAPR for 99.0%, 99.9%, and 99.99% are approximately 8.3, 10.3, and 11.8 dB, respectively. Since the OFDM signal has a high PAPR, it could be clipped in the transmitter PA, because of its limited dynamic range or nonlinearity. Higher output back-off is required to prevent performance degradation and intermodulation products spilling into adjacent channels. Therefore, RF PAs should be operated in a very large linear region. Otherwise, the signal peaks leak into nonlinear region of the PA causing signal distortion. This signal distortion introduces intermodulation among the subcarriers and OOB emission. Thus the PA should be operated with large power back-offs. On the other hand, this leads to very inefficient amplification and an expensive transmitter. Thus, it is highly desirable to reduce the PAPR. In addition to inefficient operation of the PA, a high PAPR requires larger dynamic range for the receiver analog to digital converter (ADC). To reduce the PAPR, several techniques have been proposed and used such as clipping,[9] channel coding, temporal windowing, tone reservation,[10] and tone injection. However, most of these methods are unable to achieve simultaneously a large reduction in PAPR with low complexity and without performance degradation. The PAPR ξ of an OFDM signal is defined as follows:

$$\xi = \left. \frac{\max \left| x(t) \right|^2}{E\{\left| x(t) \right|^2\}} \right|_{nT_u \le t \le (n+1)T_u}$$

In the above equation $E\{.\}$ denotes the expectation operator and n is an integer. From the central-limit theorem, for large values of N_{FFT}, the real and imaginary values of OFDM signal $x(t)$ would have Gaussian distribution. Consequently, the amplitude of the OFDM signal has a Rayleigh distribution with zero mean and a variance of N_{FFT} times the variance of one complex sinusoid. Assuming the samples to be mutually uncorrelated, the CDF for the peak power per OFDM symbol is given by

$$P(\xi > \gamma) = \left[1 - (1 - e^{-\gamma})^{N_{FFT}}\right]$$

From the above equation it can be seen that large PAPR occurs only infrequently due to relatively large values of N_{FFT} used in practice.

---

[8] The CDF of the real-valued random variable X is defined as $x \to F_X(x) = P(X \le x), \forall x \in \mathbb{R}$, where the right-hand side represents the probability that random variable X takes on a real value less than or equal to x. The CDF of X can be defined in terms of the probability density function $f(x)$ as $F(x) = \int_{-\infty}^{x} f(x)dx$. The complementary CDF (CCDF), on the other hand, is defined as $P(X > x) = 1 - F_X(x)$.

[9] Since the OFDM signal has a high PAPR, it may be clipped in the transmitter power amplifier, because of its limited dynamic range or nonlinearity. Higher output back-off is required to prevent BER degradation and intermodulation products spilling into adjacent channels. However, clipping of an OFDM signal has similar effect as impulse interference against which an OFDM system is inherently robust. Computer simulations show that for a coded OFDM system, clipping of 0.5% of the time results in a BER degradation of 0.2 dB. At 0.1% clipping, the degradation is less than 0.1 dB.

[10] In tone reservation method, the transmitter and the receiver reserve a subset of tones or subcarriers for generating PAPR reduction signals. Those reserved tones are not used for data transmission.

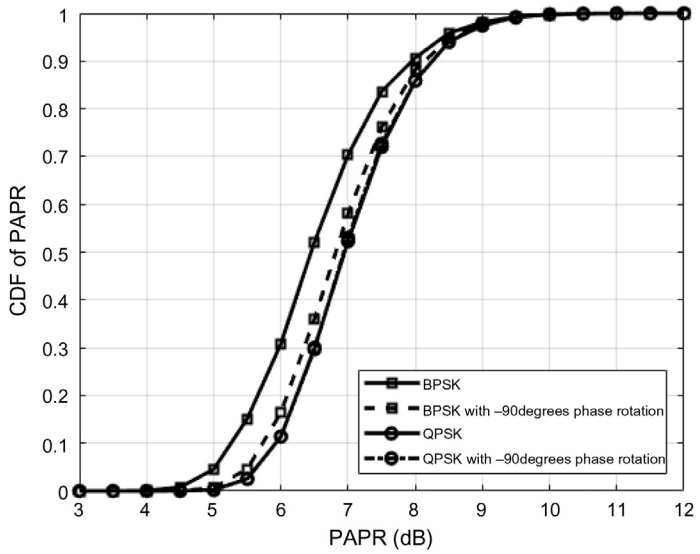


Figure 3.12
CDF of OFDM PAPR with BPSK/QPSK modulation and $N_{FFT} = 128$ [30].

Fig. 3.12 shows the CDF of OFDM PAPR for BPSK and QPSK modulation assuming a 40 MHz channel bandwidth and $N_{FFT} = 128$. We further applied 90 degrees phase rotation to the subcarriers in the upper 20 MHz of the channel and investigated the effect on the PAPR reduction [47]. It is shown that large PAPR values are less likely to occur with large FFT sizes as suggested earlier.

The PAs have generally a nonlinear amplitude response, where the output power is saturated for large input signals. Most applications require operation in the linear region of the PA where the output power is a linear function of the input. The larger the linear operation region or alternatively the higher saturation point, the more expensive the PA. Therefore, it is imperative to reduce the PAPR of the OFDM signal before processing through the PA. The wider bandwidth of NR compared to LTE increases the PAPR of the transmitted signal and makes it harder to achieve the same power efficiency as an LTE frontend. Current estimates suggest that for a single PA supplying an average transmit power of 23 dBm at the antenna, the power from the battery will be around 2.5 W, compared to around 1.8 W for current LTE UEs. The PAPR of the 5G NR signal is 3 dB higher than an equivalent LTE waveform [48], resulting in larger back-off or higher average transmit power. Another interesting observation is that for 5G CP-OFDM waveform using different modulations, there is no significant difference in the CCDF function- meaning higher order modulation has minimal impact on maximum power reduction (MPR) and PA back-off. Fig. 3.13 shows the

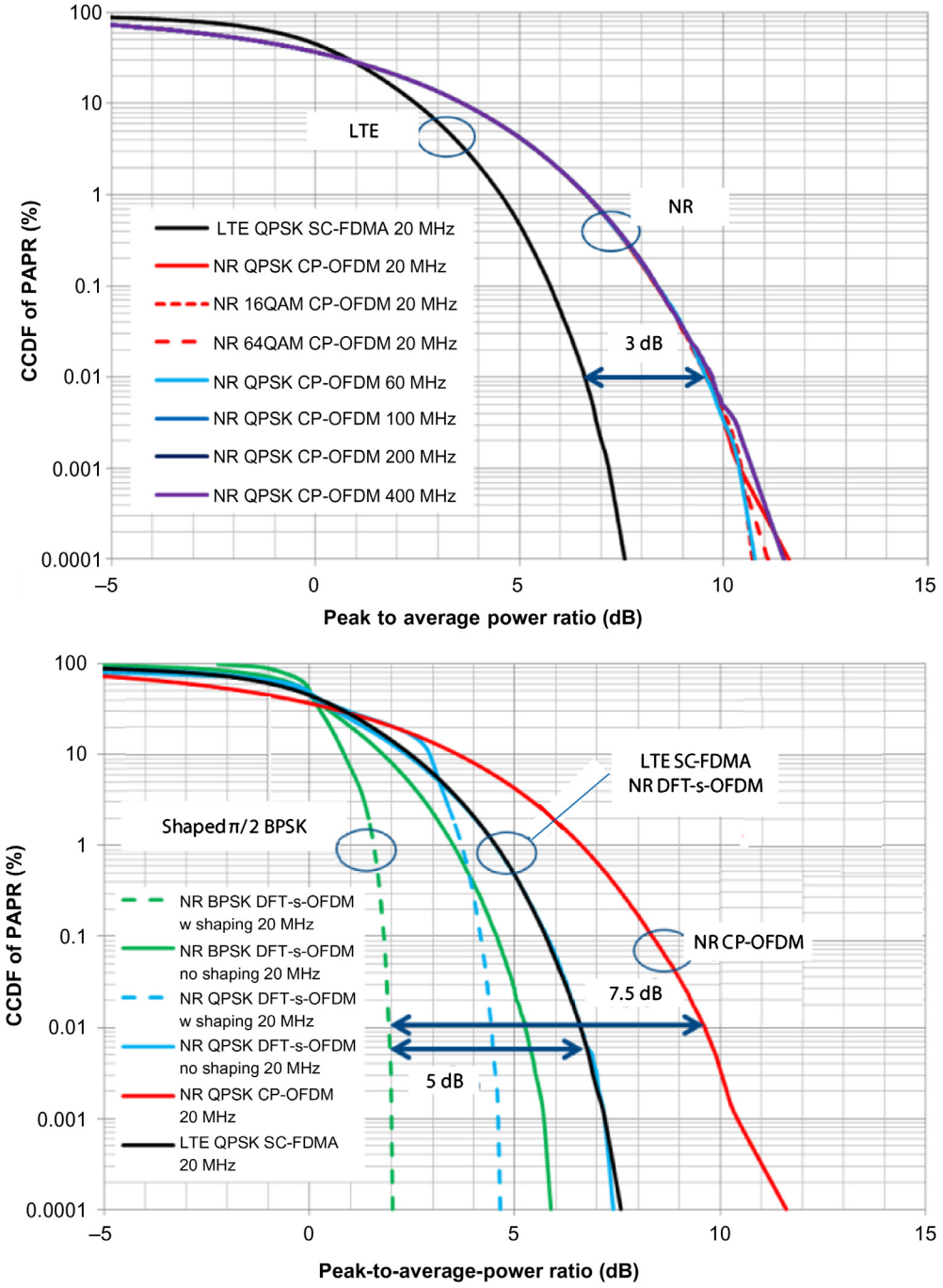


Figure 3.13
Comparison of CCDF of CP-OFDM and SC-FDMA PAPRs [48].

CCDF curves of some lower PAPR options, which can be used in cell-edge areas as well as mmWave frequency bands. We can observe that DFT-spread OFDM (DFT-S-OFDM) QPSK waveform in uplink exhibits very similar PAPR as the existing LTE single-carrier frequency division multiple access (SC-FDMA) used in the uplink; however, spectrally shaped $\pi/2-$BPSK modulation can provide up to 7.5 dB PAPR reduction which can be used in sub-6 GHz and further in mmWave bands to improve uplink link budget and coverage [48].

3.2.1.4 Error Vector Magnitude

The modulation accuracy or the permissible signal constellation fuzziness is often measured in terms of error vector magnitude (EVM) metric. In general, the EVM is defined as the square root of the ratio of the mean error vector power to the mean reference-signal power expressed as a percentage. In other words, the EVM defines the average constellation error with respect to the farthest constellation point (i.e., the distance between the reference-signal and measured signal points in I−Q plane).

In NR, the EVM measurement is conducted for all bandwidths and each NR carrier over all allocated resource blocks and downlink subframes within 10 ms measurement period. The boundaries of the EVM measurement periods are not necessarily aligned with radio frame boundaries. 3GPP defines the reference points at which the [transmitter] EVM is measured at the receiver based on which the EVM must be measured after the FFT and a zero-forcing equalizer (per subcarrier amplitude/phase correction) in the receiver [9]. The basic unit of EVM measurement is defined over one subframe in the time domain and N_{BW}^{RB} subcarriers (180 kHz) in the frequency domain as follows [9]:

$$\mathrm{EVM}=\sqrt{\frac{\sum_{t\in T}\sum_{f\in F(t)}\left|Z'(t,f)-I(t,f)\right|^2}{\sum_{t\in T}\sum_{f\in F(t)}\left|I(t,f)\right|^2}}$$

where T is the set of symbols with the considered modulation scheme being active within the subframe, $F(t)$ is the set of subcarriers within the N_{BW}^{RB} subcarriers with the considered modulation scheme being active in symbol t, $I(t,f)$ is the ideal signal reconstructed by the measurement equipment according to the relevant transmitter model, and $Z'(t,f)$ is the modified signal under test defined as follows [9]:

$$Z'(t,f)=\frac{\mathrm{FFT}\left\{z(v-\Delta\tilde{t})e^{-j2\pi\Delta\tilde{f}v}\right\}e^{j2\pi f\Delta\tilde{t}}}{\tilde{a}(f)e^{j\tilde{\phi}(f)}}$$

where $z(v)$ is the time-domain samples of the signal under test, $\Delta\tilde{t}$ is the sample timing difference between the FFT processing window relative to the nominal timing of the ideal signal, $\Delta\tilde{f}$ is the RF frequency offset, $\tilde{\phi}(f)$ is the phase response of the transmitter chain, and

$\tilde{a}(f)$ is the amplitude response of the transmitter chain. In the above equations, the basic unit of measurement is one subframe and the equalizer is calculated over 10 subframes to reduce the impact of noise on the reference symbols. The boundaries of the 10 subframes measurement periods are not necessarily aligned with radio frame boundaries.

The EVM is averaged over all allocated downlink resource blocks with the considered modulation scheme in the frequency domain, and a minimum of 10 downlink subframes. For FDD systems, the averaging in the time domain equals the 10 subframe duration of the 10 subframes measurement period from the equalizer estimation step, whereas for TDD systems, the averaging in the time domain can be calculated from subframes of different frames and should have a minimum of 10 subframes averaging length.

$$\overline{\mathrm{EVM}_{frame}} = \left[\frac{\sum_{i=1}^{N_{dl}} \sum_{j=1}^{N_i} \mathrm{EVM}_{i,j}^2}{\sum_{i=1}^{N_{dl}} N_i}\right]^{1/2}$$

where N_i is the number of resource blocks with the considered modulation scheme in subframe i and N_{dl} is the number of allocated downlink subframes in one frame. While the above expressions for calculation of the EVM are the same for FR1 and FR2, the parameters are differently defined for the two frequency ranges [9].

The permissible EVM value can be estimated from the transmitter implementation margin, if the error vector is considered noise, which is added to the channel noise. The implementation margin is the excess power needed to maintain the carrier to noise ratio intact, when going from an ideal to a realistic transmitter design. The EVM cannot be measured at the antenna connector but should be measured by an ideal receiver with certain carrier recovery loop bandwidth specified in percent of the symbol rate. The measured EVM includes the effects of the transmitter filter accuracy, DAC, modulator imbalances, untracked phase noise, and PA nonlinearity. As mentioned earlier, the error vector magnitude is a measure of the difference between the reference waveform and the measured [transmitted] waveform. In practice, before calculating the EVM, the measured waveform is corrected by the sample timing offset and RF frequency offset, then the IQ origin offset is removed from the measured waveform. The measured waveform is further modified by selecting the absolute phase and absolute amplitude of the transmitter chain.

3.2.1.5 Carrier Frequency Offset

An OFDM system transmits information as a series of OFDM symbols. The time-domain samples $x_m(n)$ of the mth OFDM symbol are generated by performing IDFT on the information symbols $s_m(k)|_{k=0,1,\ldots,N_{FFT}-1}$, as follows [30]:

$$x_m(n) = \frac{1}{N_{FFT}} \sum_{k=0}^{N_{FFT}-1} s_m(k) e^{j2\pi k(n-N_{CP})/N_{FFT}}; \quad \forall 0 \le n \le N_{FFT} + N_{CP} - 1$$

where N_{FFT} and N_{CP} denote the number of data samples and CP samples, respectively. The OFDM symbol $x_m(n)$ is transmitted through a channel $h_m(n)$ and is perturbed by a Gaussian noise $z_m(n)$. The channel $h_m(n)$ is assumed to be block-stationary, that is, it is time-invariant over each OFDM symbol. With this assumption, the output $y_m(n)$ of the channel can be represented as $y_m(n) = h_m(n)^* x_m(n) + z_m(n)$, where $h_m(n)^* x_m(n) = \sum_{k=-\infty}^{+\infty} h_m(k) x_m(k-n)$ and $z_m(n)$ is a zero-mean additive white Gaussian noise (AWGN) with variance σ_z^2. Since the channel impulse response $h_m(n)$ is assumed to be block-stationary, the channel response does not change over each OFDM symbol; however, the channel response $h_m(n)$ may vary across different OFDM symbols; thus it is a function of the OFDM symbol index m.

When the receiver oscillator is not perfectly synchronized to the transmitter oscillator, there can be a carrier frequency offset $\Delta f_{CFO} = f_{TX} - f_{RX}$ between the transmitter carrier frequency f_{TX} and the receiver carrier frequency f_{RX}. Furthermore, there may be a phase offset θ_0 between the transmitter carrier and the receiver carrier. The mth received symbol $y_m(n)$ can be represented as $y_m(n) = [h_m(n)^* x_m(n)]\exp(j\{2\pi\Delta f_{CFO}[n + m(N_{FFT} + N_{CP})]T_s + \theta_0\}) + z_m(n)$ where T_s is the sampling period. The carrier frequency offset Δf_{CFO} can be represented relative to the subcarrier bandwidth $1/(N_{FFT}T_s)$ by defining the relative frequency offset $\delta_{CFO} = \Delta f_{CFO} N_{FFT} T_s$. The carrier frequency offset attenuates the desired signal and introduces ICI, thus decreasing the SNR. The SNR of the kth subcarrier can be expressed as $\mathrm{SNR}_k(\delta_{CFO}) = \phi^2_{N_{FFT}}(\delta_{CFO}) P_h \sigma_x^2 / \left([1 - \phi^2_{N_{FFT}}(\delta_{CFO})] P_h \sigma_x^2 + \sigma_z^2 \right)$ where $\phi_{N_{FFT}}\left(\delta_{CFO}\right) = \sin(\pi\delta_{CFO}) / \left[N_{FFT} sin(\pi\delta_{CFO}/N_{FFT}) \right]$ in order to demonstrate the dependence of the SNR on the frequency offset. In the latter equation, $P_h, \sigma_x^2,$ and σ_z^2 denote the total average power of channel impulse response, variance of the signal, and variance of the additive noise, respectively. The subcarrier index k is dropped since the SNR is the same for all subcarriers. From this SNR expression, it is clearly seen that the effect of the frequency offset is to decrease the signal power by $\phi^2_{N_{FFT}}(\delta_{CFO})$ and to convert the decreased power to interference power. The SNR depends not only on the frequency offset δ_{CFO}, but also on the number of subcarriers; however, as N_{FFT} increases, $\phi^2_{N_{FFT}}(\delta_{CFO})$ converges to $\mathrm{sinc}^2(\delta_{CFO})$. Therefore, the SNR converges to $\mathrm{SNR}(\delta_{CFO}) = \mathrm{sinc}^2(\delta_{CFO}) P_h \sigma_x^2 / ([1 - \mathrm{sinc}^2(\delta_{CFO})] P_h \sigma_x^2 + \sigma_z^2)$ as N_{FFT} becomes increasingly large. In the above equations, the power of inter-carrier interference as a function of relative carrier frequency offset is defined as $P_{ICI}(\delta_{CFO}) = [1 - \mathrm{sinc}^2(\delta_{CFO})] P_h \sigma_x^2$. In practice, the subcarrier spacing is not the same among different subcarriers due to mismatched oscillators (i.e., frequency offset), Doppler shift, and timing synchronization errors, resulting in inter-carrier interference and loss of orthogonality. It can be seen that the ICI increases with the increase of the OFDM symbol duration (or alternatively decrease of subcarrier spacing) and the frequency offset. The effects of timing offset are typically less than that of the frequency offset, provided that

the CP is sufficiently large. It can be shown that the ICI power can be calculated as a function of generic Doppler power spectrum $\Lambda(\nu)$ as follows [30]:

$$P_{ICI} = \int_{-\nu_{\max}}^{\nu_{\max}} \Lambda(\nu)\left[1 - \mathrm{sinc}^2(T_s\nu)\right] d\nu$$

where $\nu_{\max}$ denotes the maximum Doppler frequency. We further assume that the transmitted signal power is normalized. It can be noted that the ICI generated as a result of carrier frequency offset is a special case of the above equation when $\Lambda(f) = \delta(f - f_{CFO})$ in which $\delta(f)$ represents the Dirac delta function. Using classic Jakes' model of Doppler spread where the spaced-time correlation function is defined as $\Phi(t) = J_0(2\pi\nu t)$ in which $J_0(x)$ denotes the zeroth-order Bessel function of the first kind, the ICI power can be written as follows:

$$P_{ICI_{Jakes}} = 1 - 2\int_0^1 (1 - f)J_0(2\pi\nu_{\max}T_s f)df$$

which approximately gives an upper bound on the ICI power due to Doppler spread. Comparison of the power of ICI generated by carrier frequency offset and Doppler spread suggests that the ICI impairment due to the former is higher than the latter.

3.2.1.6 Phase Noise

Oscillators are used in typical radio circuits to drive the mixer used for the up-conversion/down-conversion of the band-pass signal transmission. Ideally, the spectrum of the oscillator is expected to have an impulse at the frequency of oscillation with no frequency components elsewhere. However, the spectrum of a practical oscillator's output does have random variation around the oscillation frequency due to phase noise. The impact of local oscillator phase noise on the performance of an OFDM system has been extensively studied in the literature [30]. It has been shown that phase noise may have significant effects on OFDM signals with small subcarrier spacing (i.e., large OFDM symbol duration in time). Long symbol duration is required for implementing a long guard interval that can mitigate long multipath delay in single-frequency operation without excessive reduction of data throughput. The studies suggest that phase noise in OFDM systems can result in two effects: a common subcarrier phase rotation on all the subcarriers and a thermal-noise-like subcarrier de-orthogonality.

The common phase error, that is, constellation rotation, on all the demodulated subcarriers, is caused by the phase noise spectrum from DC (zero frequency) up to the frequency of subcarrier spacing. This low-pass effect is due to the long integration time of the OFDM symbol duration. This phase error can in principle be corrected by using pilots within the same symbol (in-band pilots). The phase error causes subcarrier constellation blurring rather than rotation. It results from the phase noise spectrum contained within the system bandwidth. This part of the phase noise is more crucial, since it cannot be easily corrected. The SNR

degradation caused by the common phase error can be quantified as $\mathrm{SNR}_{phase-rotation} = [I(\alpha)\beta\Delta f]^{-1}$ where Δf is the subcarrier spacing, β denotes the upper bound of the phase noise spectral mask, α is the ratio of the equivalent spectrum mask noise bandwidth and the subcarrier spacing, and $I(\alpha) = \int_{-\alpha}^{+\alpha} \mathrm{sinc}(\pi x)dx$ with $I(0.5) = 0.774, I(1) = 0.903$, and $I(\infty) = 0.774$. It can be seen that, when $\alpha > 1$, the common phase error decreases as the subcarrier spacing decreases.

To mathematically model the effect of the phase noise, let's consider the noisy output of an oscillator which contains phase noise $\varphi(t)$ as follows

$$v(t) = A_c \cos(\omega_c t + \varphi(t))$$

Let's further assume that the stochastic variation of the phase can be modeled as the output of a system with a step function impulse response and input perturbation $n(t)$ as follows:

$$\varphi(t) = \int_{-\infty}^{t} n(\tau)d\tau$$

Based on the above assumption, the single-sided power spectral density (PSD) of the phase can be written as $S_\varphi(f) = S_n(f)/(2\pi f)^2$ in which $S_n(f)$ denotes the noise PSD function. As an example, if $S_n(f)$ is modeled as white noise, then $S_\varphi(f) \approx f^{-2}$ and if $S_n(f)$ is modeled as flicker noise, then $S_\varphi(f) \approx f^{-3}$. Considering that the PSD of the phase is difficult to observe, one may alternatively look at the PSD of the oscillator's noisy output $v(t)$. It can be shown that the PSD of $v(t)$ can be calculated as follows [30]:

$$S_v(f) = \sum_{k=-\infty}^{k=+\infty} \left(\frac{a_k^2 + b_k^2}{2}\right) \frac{\beta k^2 f_c^2}{\beta^2\pi^2 k^4 f_c^4 + (f - kf_c)^2}$$

where $\{a_k\}$ and $\{b_k\}$ denote the Fourier series coefficients of $v(t)$ and β is a constant. Given that we are only interested in evaluating $S_v(f)$ at f_c, the above equation can be simplified as follows:

$$S_v(f) = \left(\frac{a_1^2 + b_1^2}{2}\right) \frac{\beta f_c^2}{\beta^2\pi^2 f_c^4 + (f - f_c)^2}$$

The above function is a Lorentz distribution.[11] We now define function $\Omega(f)$ as the ratio of noise power in 1 Hz bandwidth at offset f from center frequency to carrier power which is expressed in dBc/Hz. As theoretically expected, having a higher phase noise in the signal

[11] The Cauchy distribution is a continuous probability distribution which is also known as Lorentz distribution or Cauchy–Lorentz distribution, or Lorentzian function. It describes the distribution of a random variable that is the ratio of two independent standard normal random variables, with the probability density function $f(x; 0, 1) = \left[\pi(1+x^2)\right]^{-1}$.

does not increase the total power. A signal with higher phase noise will have smaller power near f_c and will have a broader spectrum around the center frequency. Conversely, a signal with lower phase noise has a sharper peak at the center frequency with less deviation. Therefore $\Omega(f)$ can be expressed as follows:

$$\Omega(f) = \frac{1}{\pi} \frac{\gamma}{\gamma^2 + (f - f_c)^2}, \quad \gamma = \beta \pi f_c^2$$

It can be shown that $\int_{-\infty}^{+\infty} \Omega(f) df = 1$. For higher values of β, the spectrum becomes wider with smaller magnitude of the main lobe of the spectrum. Note that a wider main lobe does not increase the total power of the carrier.

3.2.2 DFT-S-OFDM Basics and Transmission Characteristics

The LTE uplink uses SC-FDMA with CP in order to achieve inter-user orthogonality and to enable efficient frequency-domain equalization at the receiver. The DFT-spread-OFDM (DFT-S-OFDM) is a form of the single-carrier transmission technique, where the signal is generated in frequency domain, similar to OFDM as illustrated in Fig. 3.14, where the common processing blocks in OFDM and DFT-S-OFDM are distinguished from those that are specific to DFT-S-OFDM. This allows for a relatively high degree of commonality with the downlink OFDM baseband processing using the same parameters, for example, clock frequency, subcarrier spacing, FFT/IFFT size. The use of DFT-S-OFDM in the LTE uplink

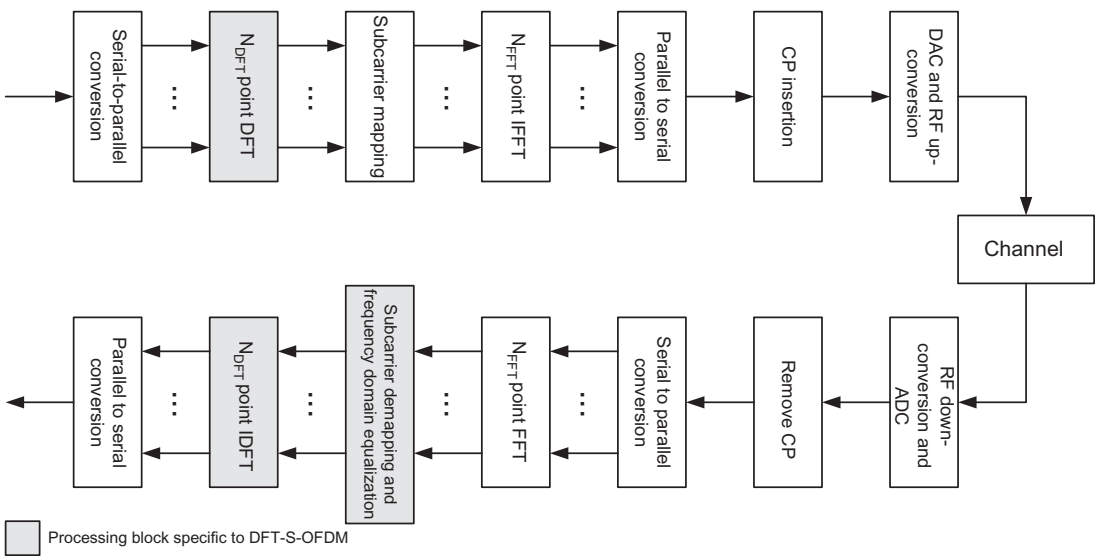


Figure 3.14
Transmitter structure for DFT-S-OFDM with localized subcarrier mapping schemes (note that $N_{DFT} < N_{FFT}$) [47].

was mainly due to relatively inferior PAPR properties of OFDM that resulted in worse uplink coverage compared to DFT-S-OFDM. The PAPR characteristics are important for cost-effective design of UE's PAs.

The principles of DFT-S-OFDM signal processing can be explained as follows. The *i*th transmitted symbol in a DFT-S-OFDM system without CP in single transmit/receive antenna case can be expressed as a vector of length N_{FFT} samples defined by $\mathbf{y} = \mathbf{F\Theta Dx}$ where $\mathbf{x} = (x_1, x_2, \ldots, x_M)^T$ is an $N_{DFT} \times 1$ vector with N_{DFT} QAM-modulated symbols (the superscript "T" denotes matrix transpose operation), $\mathbf{D}$ is an $N_{DFT} \times N_{DFT}$ matrix which performs N_{DFT}-point DFT operation, $\mathbf{\Theta}$ is the $N_{FFT} \times N_{DFT}$ mapping matrix for subcarrier assignment, and $\mathbf{F}$ performs N_{FFT}-point IFFT operation. After the propagation through the multipath fading channel and addition of the AWGN and removing the CP and going through the N_{FFT}-point FFT module, the received signal vector in the frequency domain can be expressed as $\mathbf{z} = \mathbf{F^{-1}HF\Theta Dx} + \mathbf{w}$ where $\mathbf{H}$ is the diagonal matrix of channel response and $\mathbf{w}$ is the noise vector. Note that the maximum excess delay of the channel is assumed to be shorter than the CP; therefore, the ISI can be mitigated by the CP. The amplitude and phase distortion in the received signal due to the multipath channel is compensated by a frequency-domain equalizer (FDE) and the signal at the FDE output can be described as $\mathbf{v} = \mathbf{Cz}$ where $\mathbf{C} = \mathrm{diag}\left(c_1, c_2, \ldots, c_{N_{FFT}}\right)$ is the diagonal matrix of FDE coefficients.

The FDE complex coefficients can be derived using minimum mean square error (MMSE) criterion as $c_k = H_k^* / \left(|H_k|^2 + \sigma_n^2/\sigma_s^2\right)$ where k denotes the subcarrier index, σ_n^2 denotes the variance of the additive noise, and σ_s^2 is the variance of the transmitted pilot symbol. Following the subcarrier demapping function and IDFT despreading, an $N_{DFT} \times 1$ vector $\hat{\mathbf{x}}$ containing N_{DFT} QAM-modulated symbols as an estimate to the input vector $\mathbf{x}$ is obtained at the receiver. The IDFT despreading block in the receiver averages the noise over each subcarrier. A particular subcarrier may experience deep fading in a frequency-selective fading channel. The IDFT despreading averages and spreads the fading effect, which results in a noise enhancement to all the QAM symbols. Therefore the IDFT despreading makes DFT-S-OFDM more sensitive to the noise.

As shown in Fig. 3.14, the modulation symbols in blocks of N_{DFT} symbols are processed through an N_{DFT}-point DFT processor, where N_{DFT} denotes the number of subcarriers assigned to the transmission of the data/control block. The rationale for the use of DFT precoding is to reduce the cubic metric of the transmitted signal. From an implementation point of view, the DFT size should ideally be a power of 2. However, such a constraint would limit the scheduler flexibility in terms of the amount of resources that can be assigned for an uplink transmission. In LTE, the DFT size and the size of the resource allocation is limited to products of the integers 2, 3, and 5. For example, the DFT sizes of 60, 72, and 96 are allowed, but a DFT size of 42 is not allowed [30]. Therefore, the DFT can be implemented as a combination of relatively low-complexity radix-2, radix-3, and radix-5 FFT processing blocks. The subcarrier mapping in DFT-S-OFDM determines which part of the

spectrum is used for transmission by inserting a number of zeros (i.e., null subcarriers inserted between or around the data subcarriers) in the upper and/or lower end of the frequency region.

The goal of equalization is to compensate the effects of channel distortion due to frequency selectivity and to restore the original signal. One approach to signal equalization is in the time domain using a linear equalizer, which consists of a linear filter with an impulse response $w(t)$ operating on the received signal. By selecting different filter impulse responses, different receiver/equalizer strategies can be implemented. For example, the receiver filter can be selected to compensate the radio channel frequency selectivity. This can be achieved by configuring the receiver filter impulse response to satisfy $w(t)^*h(t) = 1$ where the operator "*" denotes linear convolution. This method of filtering is known as zero-forcing equalization, which compensates the channel frequency selectivity. However, the ZF equalization may lead to significant increase in the noise level after equalization, degrading the overall link performance. This will be the case especially when the channel has large variations in its frequency response. Another alternative is to select a filter which provides a tradeoff between signal distortion due to channel frequency selectivity and the corruption due to noise/interference, resulting in a filter impulse response that minimizes the mean squared error between the equalizer output and the transmitted signal. The linear equalizers are typically implemented as a discrete-time FIR-filter with certain number of taps. In general, the complexity of such a discrete-time equalizer increases with increasing bandwidth of the signal [30,46].

An alternative to time-domain equalization is frequency-domain equalization which can significantly reduce the complexity of linear equalization. In this method, the equalization is performed on a block of data. The received signal is transformed to frequency domain using a DFT operation. The equalization is done as a frequency-domain filtering operation, where the frequency-domain filter $W(k)$ is the DFT of the corresponding time-domain impulse response $w(n)$. The equalized frequency-domain signal is then transformed to the time domain using an inverse-DFT operator. For processing of each signal block of size $N = 2^m$, the frequency-domain equalization would include two N-point DFT/IDFT operations and N complex multiplications.

With the introduction of a CP, the channel would appear as a circular convolution over a receiver processing block of size N. Therefore, there would be no need for overlap-and-discard in the receiver processing. Furthermore, the frequency-domain filter taps can now be calculated directly from an estimate of the sampled channel frequency response. Similar to the OFDM case, the drawback of using CP in conjunction with single-carrier transmission is the overhead in terms of extra power consumption and bandwidth. One method to reduce the relative CP overhead is to increase the block size N of the FDE. However, the accuracy of block equalization requires that the channel to be approximately constant over a period of time corresponding to the size of the processing block.

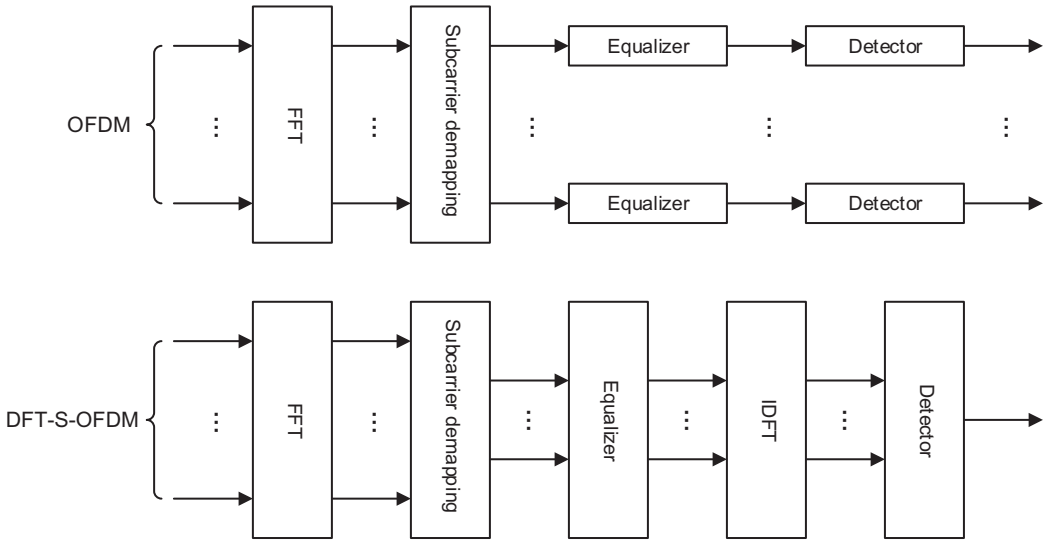


Figure 3.15
Illustration of different equalization/detection aspects of DFT-S-OFDM and OFDM [30].

The detection procedure for a DFT-S-OFDM signal is illustrated in Fig. 3.15 and is compared with that of an OFDM waveform. The transmission through a time-dispersive or equivalently a frequency-selective channel will distort the DFT-S-OFDM signal and an equalizer is needed to compensate for the effects of channel frequency selectivity. However, as shown in Fig. 3.15, a simple one-tap equalizer can be applied to each subcarrier in OFDM, whereas in the case of DFT-S-OFDM, the frequency-domain equalization function is applied to the complex-valued symbols at the output of the subcarrier demapping and prior to the IDFT operation.

It must be noted that in OFDM downlink parameterization, the DC subcarrier is unused in order to support direct conversion receiver architectures. In contrast, nulling DC subcarrier is not possible in DFT-S-OFDM since it affects the low cubic metric (CM)/PAPR property of the transmit signal. Direct conversion transmitters and receivers can introduce distortion at the carrier frequency (zero frequency or DC subcarrier in baseband) due to local oscillator leakage. In LTE downlink, this issue is avoided by inclusion of an unused DC subcarrier. However, for the uplink when using the DFT-S-OFDM waveform, the same solution may adversely impact the low CM property of the transmitted signal. In order to minimize the impact of such distortion on the packet error rate and the CM/PAPR, in LTE, the DC subcarrier of the DFT-S-OFDM signal is modulated in the same way as all the other subcarriers but the subcarriers are all frequency-shifted by half a subcarrier spacing $\Delta f/2$, resulting in an offset of 7.5 kHz relative to the DC subcarrier. Therefore two subcarriers straddle the DC location; hence, the amount of distortion affecting any individual resource block is reduced by half. In LTE, the DC subcarrier was not used because it might be subject to

disproportionally high interference due to local-oscillator leakage and intermodulation products. In fact, all LTE devices could receive the full carrier bandwidth, which was centered around the carrier frequency. The NR devices, on the other hand, may not be centered around the carrier frequency and each NR device may have its own DC located at different locations in the carrier, unlike LTE, where all devices typically have their DC coinciding with the center of the carrier. Since special handling of the DC subcarrier would have been difficult in NR, it was decided to exploit the DC subcarrier for data transmission, understanding that the quality of this subcarrier may be degraded in some conditions.

In order to demonstrate the similarities and differences between OFDM and DFT-S-OFDM processing, let's assume that one wishes to transmit a sequence of eight QPSK symbols as shown in Fig. 3.16 [30]. In the OFDM case assuming $N_{DFT} = 4$, four QPSK symbols would be processed in parallel, each of them modulating its own subcarrier at the appropriate QPSK phase. After one OFDM symbol period, a guard period or CP is inserted to mitigate the multipath effects. For DFT-S-OFDM, each symbol is transmitted sequentially. With $N_{DFT} = 4$, there are four data symbols transmitted in one DFT-S-OFDM symbol period. The higher rate data symbols require four times the bandwidth and so each data symbol occupies $N_{DFT} \times \Delta f$ Hz of spectrum assuming a subcarrier spacing of Δf Hz. After four data symbols, the CP is inserted. Note the OFDM and DFT-S-OFDM symbol periods are the same [30].

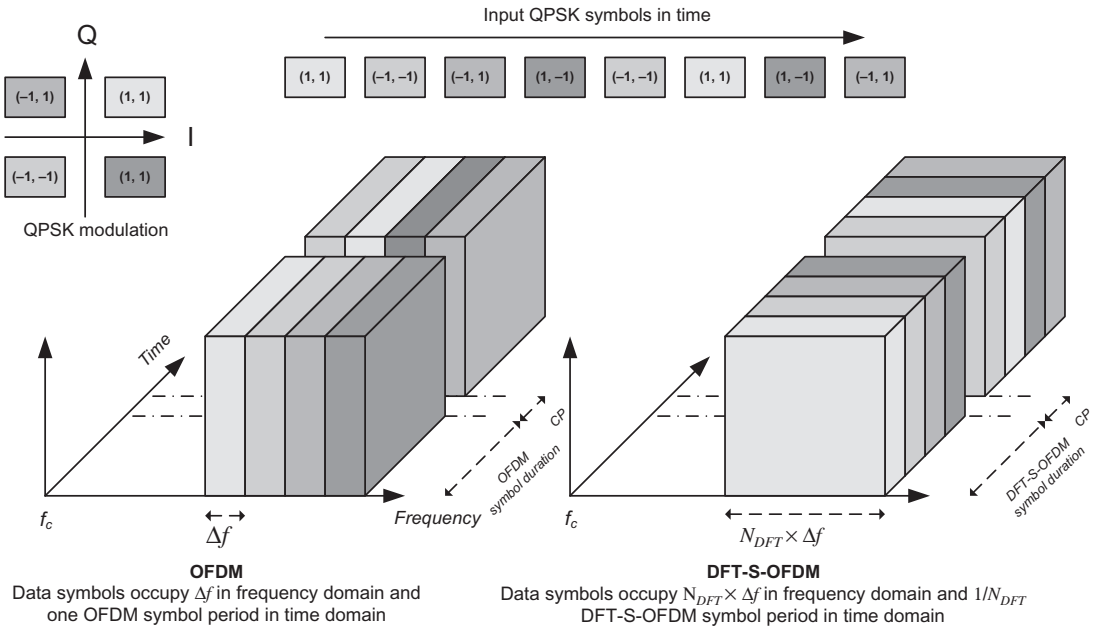


Figure 3.16
Comparison of OFDM and DFT-S-OFDM using QPSK modulation with $N_{DFT} = 4$ [30].

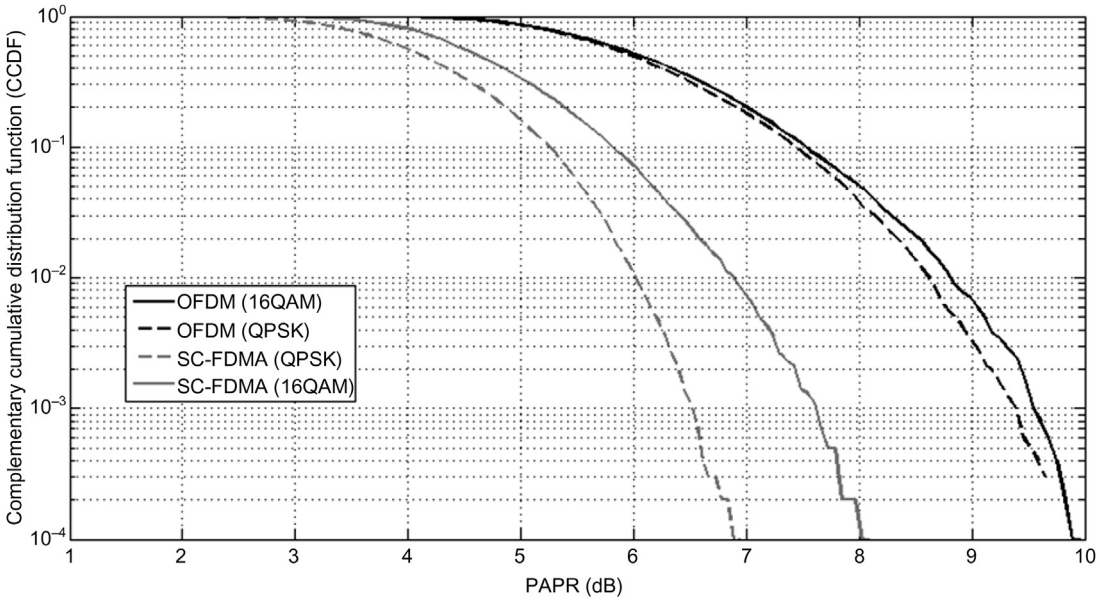


Figure 3.17
Comparison of OFDM and DFT-S-OFDM PAPRs (5 MHz bandwidth) [30,46].

As mentioned earlier, the PAPR of OFDM intrinsically is inferior to DFT-S-OFDM. Fig. 3.17 shows the comparison of complementary CDF (CCDF) of OFDM and DFT-S-OFDM PAPRs. It can be seen that the PAPR of DFT-S-OFDM is approximately 3 dB better than that of OFDM with probability of 0.99. In the case of 16QAM modulation, the PAPR of DFT-S-OFDM increases relative to that of DFT-S-OFDM with QPSK modulation, whereas in the case of OFDM, the PAPR distribution is independent of the modulation scheme because the OFDM signal is the sum of a large number of independently modulated subcarriers; thus the instantaneous power has an approximately exponential distribution, regardless of the modulation scheme applied to different subcarriers [30,46].

3.2.3 Other Waveform Candidates

In the study phase of 3GPP NR, a number of waveforms promising to improve upon OFDM waveform and to overcome the limitation of the latter were proposed and evaluated. However, when the practical aspects of implementation complexity and analog RF processing were considered, many of those candidate waveforms fail to provide any significant improvement over the status-quo, and thus 3GPP agreed to specify the OFDM as the baseline waveform for the new radio. In the following sections, we briefly describe those candidate waveforms and their respective advantages and disadvantages over OFDM.

3.2.3.1 Filtered-OFDM

To mitigate the limitations of OFDM waveform, filtered-OFDM (F-OFDM) waveform was proposed wherein subband-based splitting and filtering were used to allow independent OFDM systems operate in the assigned bandwidth. In this way, F-OFDM can overcome the drawbacks of OFDM while retaining the advantages of it. With subband-based filtering, the requirement on system-wide synchronization is relaxed and inter-subband asynchronous transmission can be supported. Furthermore, with suitably designed filters to suppress the OOB emissions, the guard band size can be reduced to a minimum. Within each subband, optimized numerology can be applied to suit the needs of certain type of services.

Fig. 3.18 shows the block diagram of a frequency-localized OFDM-based waveform. As shown in the figure, the baseband OFDM signal of each subband with its specific numerology is independently generated by processing through a spectrum shaping filter. The main purpose of this filtering is to avoid interference to the neighboring subbands. There are various approaches to the design of the spectrum shaping filter. In subcarrier filtering, the sinc(.) pulse shape of each individual subcarrier within the subband is filtered to make it more localized in frequency. An example of this method is the windowed OFDM where the subcarrier filtering is performed in the time domain by modifying the rectangular pulse shape of CP-OFDM to have smoother transitions in time at both ends. In an alternative approach known as subband filtering, the PSD of the entire subband is made well-localized without changing the CP-OFDM symbol's rectangular pulse. For this purpose, the subband CP-OFDM signal is passed through a frequency-localized filter whose bandwidth is close to the size of the subband. As a result, only a few subcarriers close to edges of the subband in frequency domain are affected by the filter, as the filter suppresses their out-of-subband side-lobes. This leads to F-OFDM signal generation. A key property of this approach is that the filter length can exceed the CP length, which allows better frequency localization than the subcarrier-based approach without causing any ISI. Although the subcarrier-based approach provides a lower complexity, it cannot achieve the frequency localization

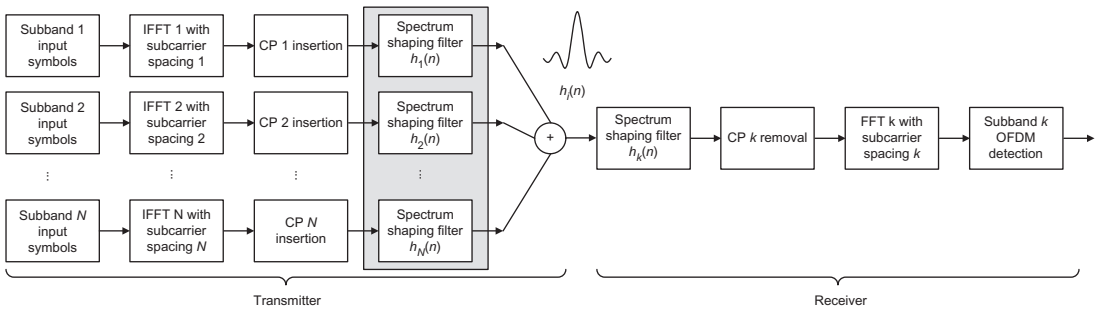


Figure 3.18
Illustration of transmit/receive processing for frequency-localized OFDM-based waveforms [49].

performance of the subband-based approach, and further causes ISI in multipath channels with large delay spread. A combination of the two approaches can provide a tradeoff between complexity and frequency localization. In particular, in the latter approach, the CP-OFDM signal of each subband is first subcarrier-filtered with an excess window length smaller than that of the original subcarrier-based approach. Then, the windowed signal is subband filtered with a filter length smaller than that of the original subband-based approach.

In the receiver side, in order to filter out the signals of the neighboring subbands, the received signal in baseband is first passed through the receiver spectrum shaping filter. Subcarrier-based, subband-based, or a composite approach can be employed by the receiver, independent of the approach employed by the transmitter. After the subband spectrum shaping, the resulting signal is processed by the regular OFDM processing within that subband.

The filters used in F-OFDM processing must satisfy a number of criteria. The passband of the filter should be as flat as possible over the subcarriers contained in the subband. This ensures that the distortion caused by the filter in the data subcarriers, especially the subband edge subcarriers, is minimal. The frequency roll-off of filter should start from the edges of the passband and the transition band of the filter should be sufficiently steep. This ensures that the system bandwidth is utilized as efficient as possible and the guard band overhead is minimized. Also, the neighboring subband signals with different numerologies can be placed next to each other in frequency with minimal number of guard subcarriers. The filter should further have sufficient stop-band attenuation to ensure that the leakage into the neighboring subbands is negligible.

The F-OFDM waveform processing introduces negligible delay at the receiver side. Signal processing delay of F-OFDM depends on the receiver processing capabilities. It should be noted that the only extra signal processing block in F-OFDM receiver compared to CP-OFDM is the receiver subband spectrum shaping filter. The delay due to spectrum shaping filter is implementation-specific and depends on the receiver processing capabilities. The rest of receiver processing blocks in F-OFDM, for example, FFT block size, channel estimation/equalization are the same as those in CP-OFDM [49]. The F-OFDM concept can be used in the asynchronous access of multiple UEs in the uplink as shown in Fig. 3.19.

3.2.3.2 Filter Bank Multicarrier

Filter bank multicarrier (FBMC) is an OFDM-based waveform wherein subcarriers are individually processed through filters that suppress their side-lobes, making them strictly band-limited. The transmitter and receiver may still be implemented through FFT/IFFT blocks or polyphase filter structures and band-limitedness may offer larger spectral efficiency than OFDM. During the study of waveforms, FBMC was found promising mainly due to signal band-limitedness in order to relax synchronization requirements in the uplink and/or in the

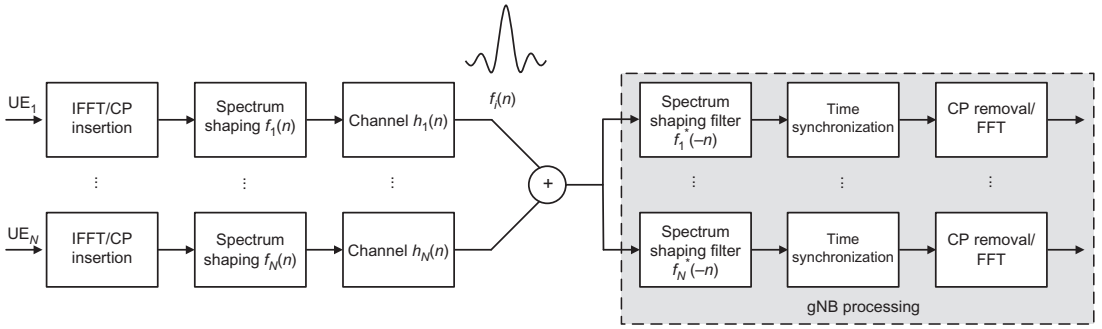


Figure 3.19
F-OFDM uplink asynchronous communication.

downlink with coordinated transmission, its greater robustness to frequency mis-alignments among users when compared to OFDM, and its more flexible exploitation of frequency white spaces in cognitive radio networks. The rectangular impulse adopted in OFDM systems is not well-localized in time and frequency, making it sensitive to timing and frequency offsets (e.g., introduced by channel, or local oscillator mismatch). As we discussed earlier, ideal time and frequency well-localized pulse does not exist in practice for the conventional OFDM according to Balian–Low theorem.[12] However, if pulse amplitude modulation (PAM) symbols instead of QAM symbols are considered, time and frequency well-localized pulse can be achieved in a multicarrier system called FBMC. The transmit signal can be expressed as

$$x(t) = \sum_{k=-\infty}^{\infty} \sum_{n=0}^{N_{\text{sub}}-1} j^{n+k} s_{k,n} g\left(t - kT/2\right) e^{j(2\pi/T)nt}$$

where $g(t)$ is a square-integrable function on real domain (Gabor set), which is manifested as the rectangular pulse in OFDM, and $s_{k,n}$ denotes real-valued data symbols.

In FBMC, the pulse $g(t)$ can be designed to achieve better time and frequency localization properties using filter design methods. Usually, the prototype filter $g(t)$ spans an integer K (overlapping factor) multiple length of symbol period $T_F = KT$. It must be noted that real

[12] In mathematics, the Balian–Low theorem in Fourier analysis states that there is no well-localized window function or Gabor function either in time or frequency domain for an exact Gabor frame. Let g denote a square-integrable function on the set of real numbers, and consider the so-called Gabor system $g_{m,n}(x) = g(x - na)e^{2\pi jmbx}$ for integers m and n, and $a, b > 0$ satisfying $ab = 1$. The Balian–Low theorem states that if $\{g_{m,n}: m, n \in \mathbf{Z}\}$ is an orthonormal basis for the Hilbert space $L^2(\mathbb{R})$, then either $\int_{-\infty}^{\infty} x^2 |g(x)|^2 dx = \infty$ or $\int_{-\infty}^{\infty} \phi^2 \left|\widehat{g}(\phi)\right|^2 d\phi = \infty$ [Note : $\widehat{g}(\phi)$ is the Fourier transform of $g(x)$]. The Balian–Low theorem has been extended to exact Gabor frames.

and imaginary data values alternate on subcarriers and symbols, which is called offset QAM (OQAM). Since PAM symbols convey only one half of information content compared to QAM symbols, a data rate loss factor 2 is implicit. Nevertheless, the symbol period in FBMC is also halved to $T/2$ in order to compensate for the efficiency loss of OQAM modulation. Furthermore, CP is not essential anymore in FBMC due to the well-localized pulse shape. Filtering can use different overlap factors (i.e., K factor) to provide varying levels of OOB rejection. As K factor is reduced, the OOB characteristics have a spectrum-rejection profile similar to that of OFDM. The critical step for FBMC design is to implement filters for each subcarrier and to align multiple filters into a filter bank. One way to implement the filter bank is to design a prototype filter. Once the prototype filter is designed, the next step is to make a copy of the prototype filter and shift it to neighboring subcarriers as illustrated in Fig. 3.20.

The comparison of the PSDs of the OFDM and FBMC signals is shown in Fig. 3.21. The FBMC signal was processed with overlapping factors $K = 2$ and 3. It is shown that with the increase of the overlapping factor, the OOB emissions of the FBMC signal significantly decreases; however, the signal processing complexity and latency prohibitively increases relative to that of the OFDM processing.

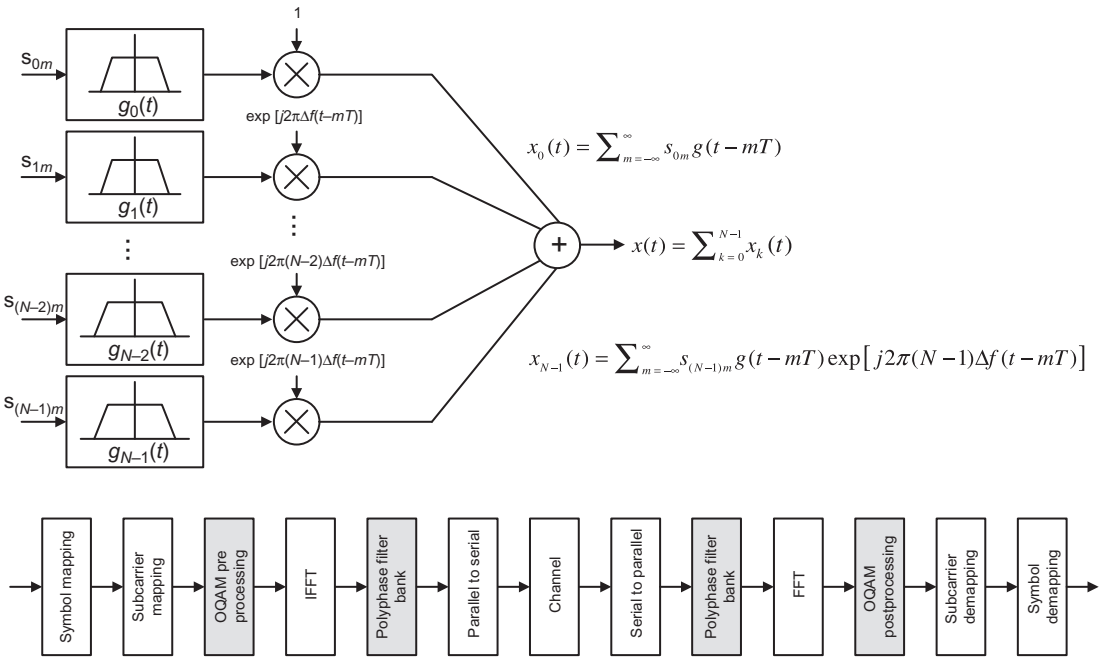


Figure 3.20
Illustration of FBMC concept and transmitter/receiver architecture [47].

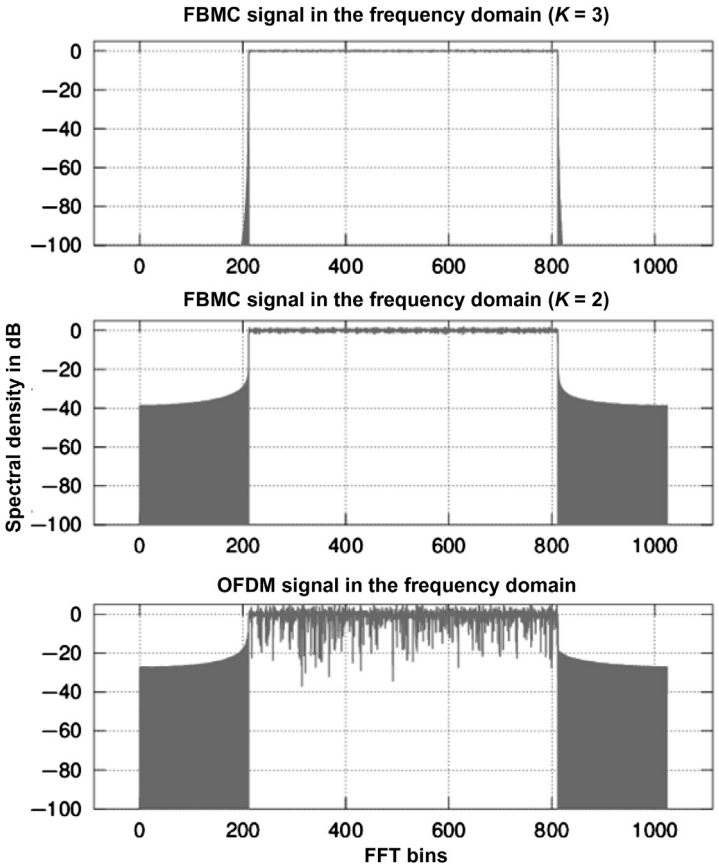


Figure 3.21
Comparison of OFDM and FBMC signals in the frequency domain [38].

3.2.3.3 Universal Filtered Multicarrier

Universal filtered multicarrier (UFMC) is a generalization of OFDM and FBMC. The ultimate goal of UFMC is to combine the advantages of OFDM and FBMC while avoiding their main drawbacks. By filtering groups of adjacent subcarriers, the side-lobe levels (compare to OFDM) and the prototype filter length (compare to FBMC) can be simultaneously and significantly reduced. The kth OFDM signal over the ith physical resource block (i.e., 12 adjacent subcarriers in NR) can be expressed as follows [38]:

$$x_{k,i}(m) = \sum_{n \in S_i} s_{k,n} e^{j(2\pi/N_{\text{sub}})kn}, \quad m = 0, \ldots, N_{\text{sub}} - 1$$

where S_i is a set which contains consecutive subcarrier indices that are assigned to the ith physical resource block. This signal is then filtered by an FIR-filter $f_i(n)$ with the length of

L_F. The UFMC scheme applies filtering on a per subband basis, reducing complexity of the baseband processing algorithms. Thus, the kth transmit symbol can be written as

$$\tilde{x}_k(m) = \sum_i \sum_{l=0}^{L_F-1} f_i(l) x_{k,i}(m-l), \quad m = 0, \ldots, N_{\text{sub}} + L_F - 2$$

The FIR-filter can be differently designed for each physical resource block. Let's assume that we use an identical Chebyshev filter with variable side-lobe attenuation for all physical resource blocks and the filter is shifted to the center frequencies of the physical resource blocks. The filter ramp-up and -down regions at the beginning and the end of individual UFMC symbols provide somewhat ISI protection, in the presence of channel delay spreads and timing offsets. With very high delay spreads, sophisticated multi-tap equalizers must be applied.

There is no time overlap between subsequent UFMC symbols. The symbol duration is $N_{sub} + L_F - 1$ with N_{sub} being the FFT size of the IFFT spreaders (the size of the subbands) and L_F the length of the filter. Similar to FBMC, in UFMC typically the FFT window size is increased, resulting in a higher implementation complexity. Also in UFMC the insertion of a guard interval or CP is optional. Another feature of the unified frame structure is the usage of multiple signal layers. The users can be separated based on their interleavers as it is done in interleave division multiple access scheme. This will introduce an additional degree of freedom for the system, improve robustness against cross-talk, and help to exploit the capacity of the multiple access channel (uplink) [45].

The comparison of the PSDs of the OFDM and UFMC signals is shown in Fig. 3.22. In the processing of the UFDM signal, a Chebyshev filter with $L_F = 74$ and side-lobe level attenuation of 40 dB has been used. Furthermore, we assume $N_{FFT} = 1024$ and $N_g = 0$. It is shown that while the relative complexity of UFMC is more manageable than FBMC, the OOB components are significantly more suppressed compared to that of the OFDM signal.

An alternative mathematical representation of the UFDM signal generation and processing can be given as follows:

$$\underbrace{\mathbf{x}[k]}_{[(N+L_F-1),1]} = \sum_{n=1}^{N_{SB}} \underbrace{\mathbf{F}_{n,k}}_{[N+L_F-1),1]} \underbrace{\mathbf{V}_{n,k}}_{[N,N_n]} \underbrace{\mathbf{S}_{n,k}}_{[N_n,1]}$$

where N, L_F, N_{SB}, and N_n denote the FFT size, the filter length, the number of subbands, and the number of complex QAM symbols, respectively; $[n, k]$ represents the subband index and user number; $\mathbf{F}_{n,k}$,$\mathbf{V}_{n,k}$, and $\mathbf{S}_{n,k}$ denote a Toeplitz matrix comprising the filter impulse response, an IDFT matrix corresponding to the subband location, and a symbol matrix, respectively. The above mathematical model is illustrated in Fig. 3.23.

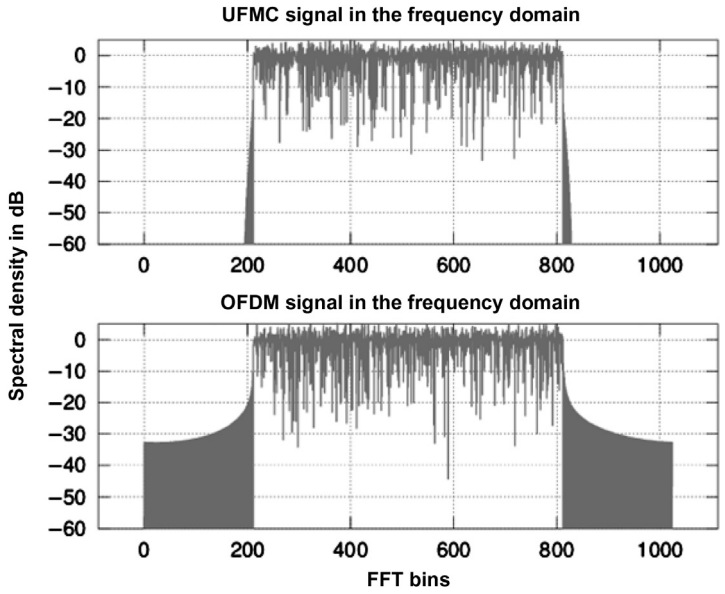


Figure 3.22
Comparison of OFDM and UFMC signals in the frequency domain [38].

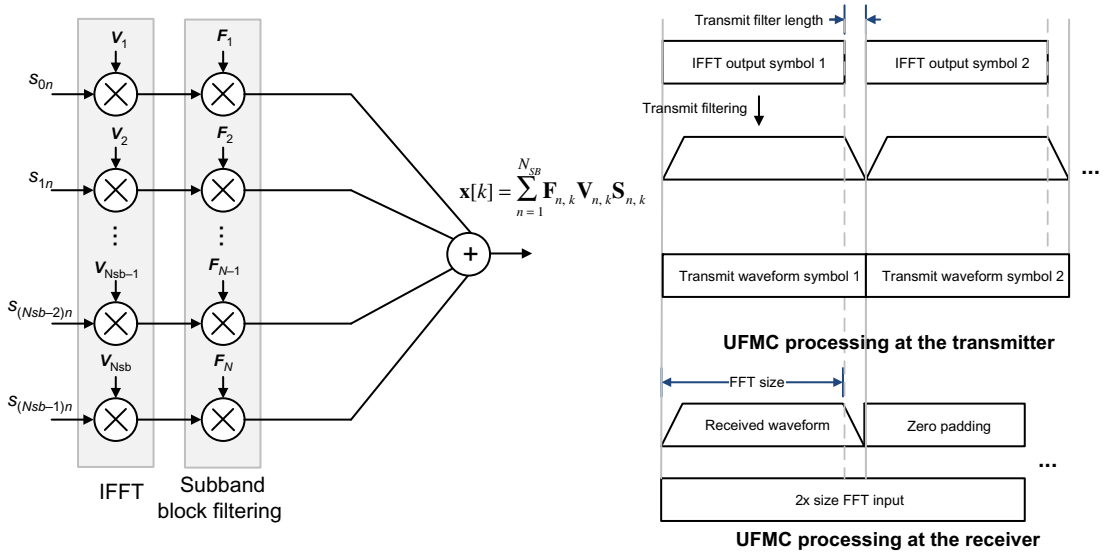


Figure 3.23
UFDM signal processing [42,47].

3.2.3.4 Generalized Frequency Division Multiplexing

Generalized frequency division multiplexing (GFDM) is a flexible multicarrier modulation scheme. The process is performed block-by-block, where each GFDM block consists of a number of K subcarriers and M sub-symbols. By setting the number of subcarriers and the number of sub-symbols to one, GFDM reduces to single-carrier frequency domain equalization and CP-OFDM as its special cases. Furthermore, pulse shaping with a prototype filter $g_{0,0}(n)$ is another flexibility in GFDM to reduce OOB emissions. In contrast to linear convolution used in FBMC, GFDM uses circular convolution. Let $g_{k,m}(n)$ denote the pulse-shaping filter corresponding to the data symbol $s_{k,m}$ that is transmitted at subcarrier m and time k. It can be shown that

$$g_{k,n}(n) = g_{0,0}[(n - mK) \bmod N] e^{j2\pi kn/K}, \quad N = M \times K$$

In the above equation N denotes the number of symbols within a GFDM block (GFDM block size). Thus, the time-domain signal $x(n)$ of a GFDM block is expressed as

$$x(n) = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} s_{k,m} g_{k,m}(n), \quad n = 0, 1, \ldots, N-1$$

A CP and a cyclic suffix can be optionally added in the GFDM data block. Furthermore, a raised-cosine filter with configurable roll-off factor β is used for filtering. The GFDM signal processing stages are illustrated in Fig. 3.24.

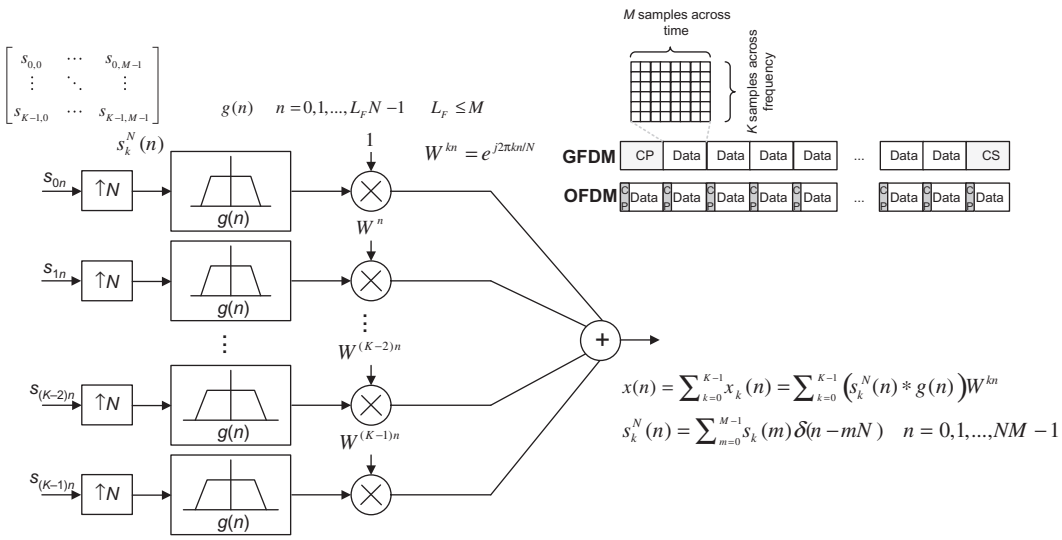


Figure 3.24
Illustration of GFDM signal processing [45,47].

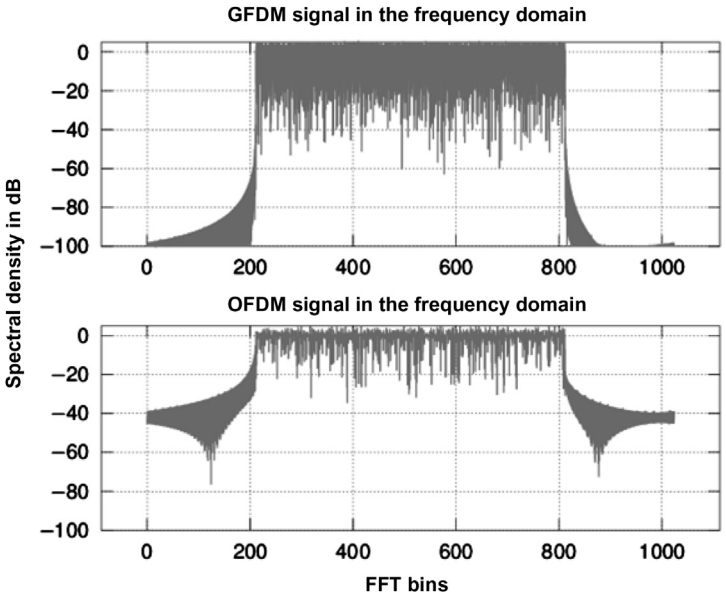


Figure 3.25
Comparison of OFDM and GFDM signals in the frequency domain [38].

In one aspect, GFDM is similar to FBMC where a prototype filter is used to suppress OOB emissions. However, for GFDM, multiple OFDM symbols are grouped into a block and a CP is added to the block. Within a block, the prototype filter is cyclic-shift in time, for different OFDM symbols. Therefore, better OOB leakage suppression can be achieved relative to CP-OFDM. However, the approach results in a complicated receiver to handle the ISI and ICI. Furthermore, the prototype filter may require more complicated modulation, for example, OQAM as in FBMC, and more complex receiver architecture. Higher block processing latency is inevitable in GFDM given that there is no possibility for pipelining, and multiplexing with CP-OFDM requires a large guard band, which would add to the overhead. The comparison of the PSDs of the OFDM and GFDM signals is shown in Fig. 3.25. In the processing of the GFDM signal, a pulse-shaping filter with roll-off factor $\beta = 0.1$ and $N = 10$ symbols were used. Furthermore, we assume that $N_{FFT} = 1024$, OFDM $N_g = 10$ and GFDM $N_g = 100$ samples. It is shown that while the relative complexity of GFDM processing is more than that of OFDM, the OOB components are significantly more suppressed compared to that of the OFDM signal.

3.2.3.5 Faster Than Nyquist Signal Processing

Faster than Nyquist (FTN) is a non-orthogonal transmission scheme which was one of the approaches initially considered for 5G systems that was expected to improve the spectrum

efficiency by increasing the data rate. It was observed a few decades ago that the binary sinc pulse can be transmitted faster than what the Nyquist theorem states without increasing the bit error rate and despite of ISI. The idea was extended to frequency domain to reduce subcarrier spacing. The transmit signal of FTN can be expressed as follows [4]:

$$x(t) = \sum_{k=-\infty}^{\infty} \sum_{n=0}^{N_{\text{sub}}-1} s_{k,n} g(t - k\Delta_T T) e^{j2\pi\Delta_F nt/T}$$

where $\Delta T \leq 1$ is the time compression factor, which means that the pulses are transmitted faster a factor of $1/\Delta T$ and $\Delta F \leq 1$ is the frequency compression factor, which means the spectral efficiency is increased by a factor of $1/\Delta F$. The FTN transmitter structure is depicted in Fig. 3.26. Since the time and frequency spacing varies for different FTN systems, direct implementation cannot provide sufficient implementation flexibility. An FTN mapper based on projection scheme has been designed and used in FTN signaling systems to provide flexibility. The FTN mapper is shown in Fig. 3.26. A cyclic extension is needed in the modulation block, with which the system can switch easily between FTN and Nyquist modes. As mentioned before, FTN signaling inevitably introduces ISI in the time domain and/or ICI in the frequency domain when its baud rate is over the Nyquist one (see Fig. 3.27). Therefore a very important issue is how to design a receiver with ISI and/or ICI suppression capability to recover the original transmitted data.

3.2.3.6 Comparison of the Candidate Waveforms

A number of candidate waveforms were studied by 3GPP for NR, before CP-OFDM was selected as the default waveform for the downlink and uplink, which were based on FFT/IFFT processing, additional filtering, windowing, or precoding, in order to achieve higher time-frequency localization and lower OOB spectral leakage, and higher throughput.

Filtering is a straightforward way to suppress OOB emissions by applying a digital filter with prespecified frequency response. Some waveforms like F-OFDM and UFDM belong to this category. However, the delay spread of the equivalent composite channel may exceed the CP size and guard period in TDD systems, which results in ISI and imposes restrictions on downlink-to-uplink switching time. Furthermore, the promised OOB emission performance may diminish significantly when PA nonlinearity and other non-ideal RF processing effects are taken into account. At the cost of increased PAPR, filtering techniques are generally known to be unfriendly to communication at high carrier frequencies.

Windowing is used to prevent steep changes across two consecutive OFDM symbols in order to reduce the OOB emissions. Multiplying the time-domain samples located in the extended symbol edges by raised-cosine window coefficients is a widely used realization as chosen by windowed OFDM and weighted overlap-and-add OFDM waveforms. This

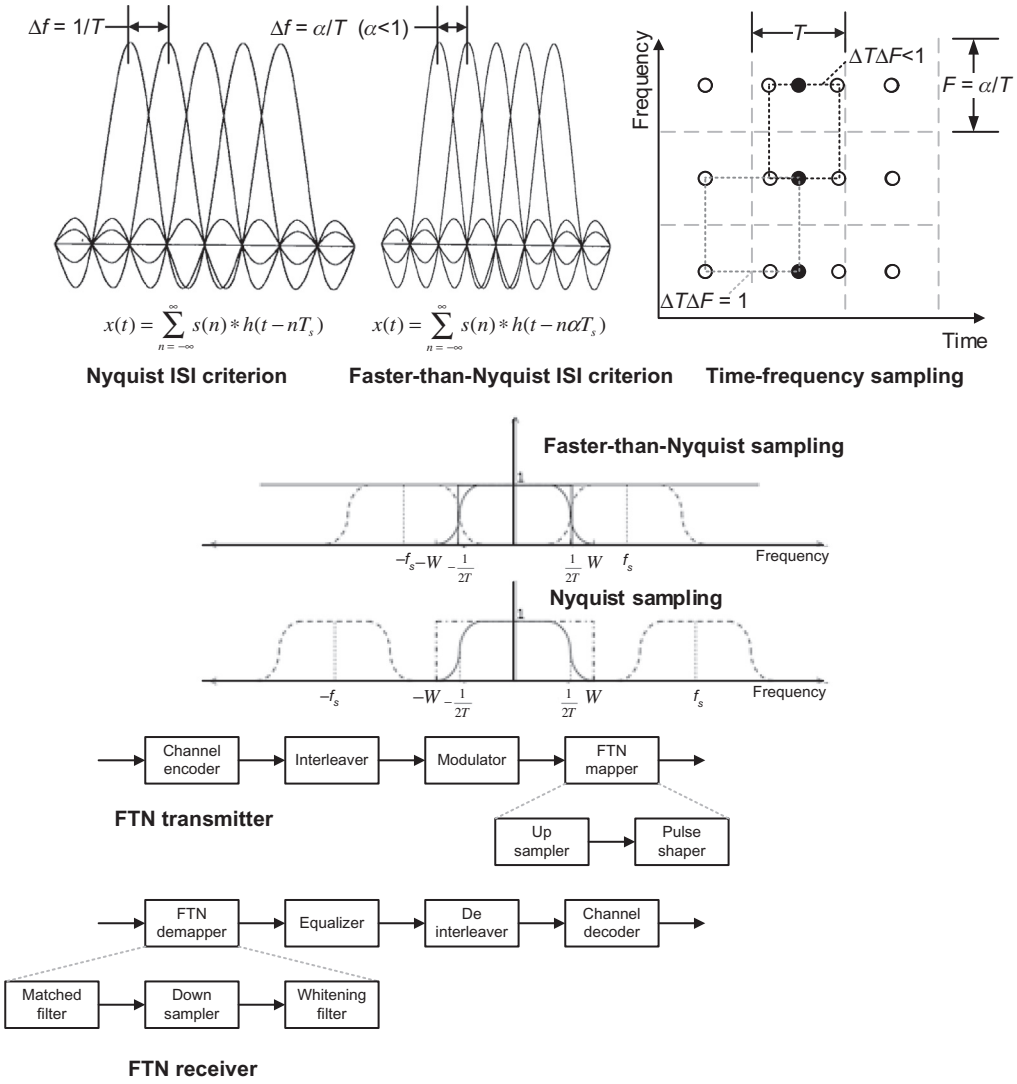


Figure 3.26
Illustration of FTN concept and transmitter and receiver architecture [44].

technique generally has little or no effect on PAPR increase and has lower complexity compared to that of filtering techniques. Nevertheless, the detection performance might be degraded because of ISI caused by symbol extension.

The linear processing of input data prior to IFFT is usually known as precoding, and may be helpful to improve OOB emissions and PAPR reduction. One example is DFT-S-OFDM

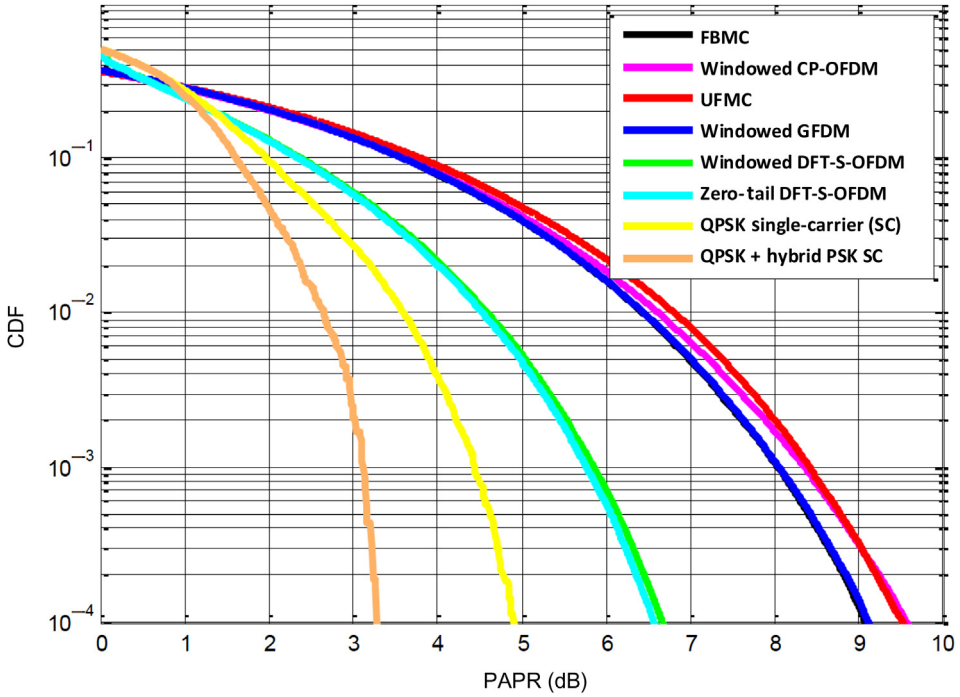


Figure 3.27
Comparison of PAPR performance of some prominent variants [47].

waveform that was adopted in LTE uplink transmission because of its low PAPR properties relative to conventional OFDM. Some variants of DFT-S-OFDM were proposed for NR such as zero-tail (ZT) DFT-S-OFDM by omitting the CP and letting the tail samples taper to zero. The DFT-S-OFDM-based waveforms, in contrast to filter-based waveforms, usually make it easier to maintain PA linear operation with less deterioration from lowering OOB emissions. Moreover, an appropriate modification of modulation schemes, such as $\pi/2$-BPSK can greatly assist such waveforms in achieving an extremely low PAPR. Note that in the absence of redundant intervals, ISI can still occur. Among DFT-based precoding techniques, other types of precoding matrices often have undesirable complexity and compatibility issues.

PAPR is one of the often-mentioned disadvantages of OFDM waveform. In practice, the crest factor reduction (CFR) techniques are applied to reduce the PAPR and digital predistortion algorithms will then correct for any distortion caused by the analog hardware used

to amplify the signal. Both techniques will allow more efficient PA design and help mitigate major limitations of PAPR and spectral regrowth.[13] Traditionally these techniques were only applied at the base station side, but currently, they are also used in mobile devices, mainly from the aspect of reducing power consumption. The use of envelope tracking[14] to reduce the static power consumption is an example of such techniques. Fig. 3.27 compares the PAPR performance of the prominent waveforms and shows that despite additional complexity and latency, the relative PAPR performance of multicarrier techniques remains about the same as OFDM and inferior to DFT-S-OFDM [47].

In the spectrum of rectangular pulse (i.e., a sinc function), besides the desired peak, there are some side-lobes that result in a theoretical infinite bandwidth of the rectangular pulse function, causing OOB emissions. Moreover, consecutive OFDM symbols are independent of each other; thus there is an inherent discontinuity in the time domain between them. In this way OFDM differs from single-carrier modulated signals after digital filtering. This discontinuity translates into spectral spurs in the frequency domain. This typical characteristic can be improved by applying time-domain windowing that smooths out the transition from one symbol to another. However, this technique introduces an overlap between consecutive symbols that impacts signal quality and results in higher EVM. The transition time defines the duration of the overlap between two symbols. For a sampling rate of 30.72 MHz (20 MHz LTE signal), a transition time of 1 μs translates into 30 samples overlap.

Fig. 3.28 shows the ideal case by means of connecting the signal generator directly to a spectrum analyzer. The idea is to demonstrate the impact of a nonlinear device on the highlighted advantage of 5G waveform candidates, where any nonlinearity will result in a spectral regrowth, and there is a risk that this spectral regrowth may undermine any optimization due to the waveform design. The improved spectral characteristics of the candidate waveforms are clearly visible in the top snapshot. In a second step, a nonlinear amplifier is introduced into the signal path. The top and bottom snapshots in Fig. 3.28 compare the LTE OFDM signal with FBMC, UFMC, and GFDM waveforms under two different conditions. A generic PA, which supports a frequency range of 50 MHz to 4 GHz, is used to demonstrate the effects of nonlinearity on the waveforms. The maximum input power for the PA is 0 dBm, and it has a typical

[13] Intermodulation products or spurs can develop within the analog and digital transmitters in combined systems using high-level injection. In some cases spurs can result in suboptimal signal quality or even cause stations to interfere with each other's signals. The term spectral regrowth is used to describe intermodulation products generated when a digital transmitter is added to an analog transmission system.

[14] Envelope tracking is an approach to RF amplifier design in which the power supply voltage applied to the RF power amplifier is continuously adjusted to ensure that the amplifier is operating at peak efficiency for power required at each instant of transmission. A conventional RF amplifier is designed with a fixed supply voltage and operates most efficiently only when operating in compression region. Amplifiers operating with a constant supply voltage become less efficient as the crest factor of the signal increases, because the amplifier spends more time operating below peak power and, therefore, spends more time operating below its maximum efficiency.

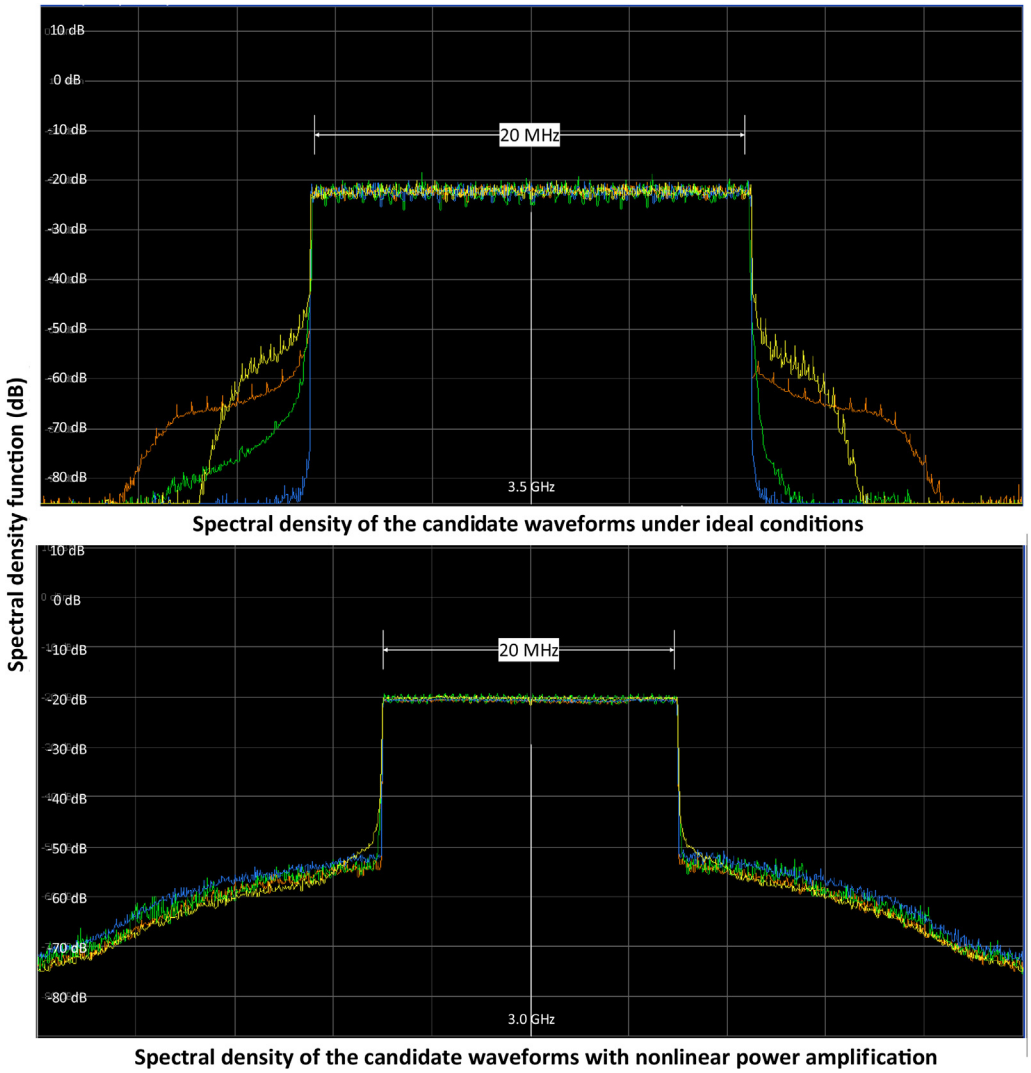


Figure 3.28
Comparison of a 20 MHz LTE downlink OFDM signal (yellow) with FBMC (blue), UFMC (green), and GFDM (orange) signals under ideal (top) and non-ideal conditions [45]. For color interpretation of this figure, please refer web version.

gain of 20 dB. Maximum achievable output power for the PA is +20 dBm. At 0 dBm input power, the PA starts to enter the saturation region. Higher input power would mean that the PA would be operating in the compression region. Fig. 3.28 further shows the result of a measurement with the same signal configuration for an input power of −2 dBm to the amplifier. The spectral advantages of the candidate waveforms seem to have almost disappeared compared to

a 20 MHz LTE downlink signal. When using typical input power of 0 dBm, the advantages of non-OFDM waveforms will completely vanish [45].

3.3 Multiple-Access Schemes

A cellular network consists of a number of fixed base stations distributed across a geographical area. The coverage area is divided into cells and a mobile station communicates with one or more base stations in its proximity. There are two main issues in the physical and medium access layers of cellular communication schemes: multiple access and interference management. The first issue addresses how the overall radio resources of the system are shared by the users in the same cell (intra-cell) and the second issue addresses the interference caused by simultaneous signal transmissions in different cells (inter-cell). At the network layer, an important issue is to provide and maintain seamless connectivity to the users as they move from one cell to another and thus switching communication link from one base station to another through an operation known as handover.

There are various multiple-access schemes that have been studied and used in wireless systems in the past decades, which allow the network to share the available radio resources (i.e., time, frequency, code, space, power) among a number of active users in the cell in the downlink and uplink. Fig. 3.29 illustrates the concept of resource sharing in some prominent multiple-access schemes. As mentioned earlier, orthogonal frequency division multiple access (OFDMA) has been a promising MA scheme that has been used in mobile broadband radio access technologies such as NR and LTE. The new radio uses a symmetric OFDMA scheme in the downlink and uplink, whereas LTE uses OFDMA and SC-FDMA as the MA schemes in the downlink and uplink, respectively.

In addition, the non-orthogonal concept can be applied to MA scenarios. Sparse code multiple access (SCMA), non-orthogonal multiple access (NOMA), and mult-iuser shared access (MUSA) are examples of non-orthogonal multiple access schemes that were studied in 3GPP Rel-16 [23,50–52]. These techniques can superimpose signals from multiple users in the code domain or the power domain to enhance the system-access performance and potentially allow asynchronous access in the uplink.

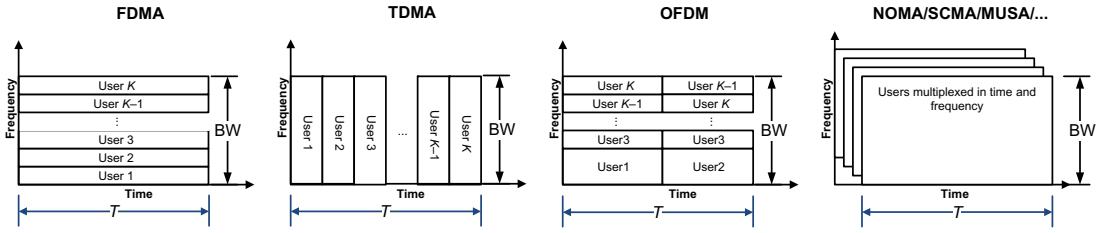


Figure 3.29
Illustration of various multiple access concepts [47].

3.3.1 Orthogonal Frequency Division Multiple Access

OFDMA is the multi-user variant of the OFDM scheme where multiple access is achieved by assigning subsets of time-frequency resources to different users, allowing simultaneous data transmission from several users. In OFDMA, the radio resources are 2D regions over time (an integer number of OFDM symbols) and frequency (a number of contiguous or non-contiguous subcarriers). Similar to OFDM, OFDMA employs multiple closely spaced subcarriers that are divided into groups of subcarriers where each group is called a resource block. The grouping of subcarriers into groups of resource blocks is referred to as sub-channelization. The subcarriers that form a resource block do not need to be physically adjacent. In the downlink, a resource block may be allocated to different users. In the uplink, a user may be assigned to one or more resource blocks. Sub-channelization defines subchannels that can be allocated to mobile stations depending on their channel conditions and service requirements. Using sub-channelization, within the same time slot (i.e., an integer number of OFDM symbols) an OFDMA system can allocate more transmit power to user devices with lower SNR and less power to user devices with higher SNR. Sub-channelization also enables the base station to allocate higher power to sub-channels assigned to indoor mobile terminals resulting in better indoor coverage. In OFDMA, an OFDM symbol is constructed of subcarriers, the number of which is determined by the FFT size. There are several subcarrier types: (1) data subcarriers are used for data transmission, (2) pilot or reference-signal subcarriers are utilized for channel estimation and coherent detection, and (3) null subcarriers that are not used for pilot/data transmission. The null subcarriers including the DC subcarrier (if it exists) are used for guard bands. The number of used (or occupied) subcarriers is always less than the FFT size. The guard bands are used to allow spectrum sharing and to reduce the adjacent channel interference and OOB emissions. The sampling frequency is selected to be greater than or equal the channel bandwidth. The number of time samples in a radio frame is always an integer and to further simplify the design of analog transmit filter, the sampling frequency is scaled by a factor greater than one (e.g., in LTE, the sampling frequency for 20 MHz bandwidth is 30.72 MHz).

In order to explain the signal processing concepts involved in an OFDMA transmission system, we use the generic transmitter model that is illustrated in Fig. 3.30, which shows the baseband structure of a general multicarrier transmitter that is applicable to a variety of multicarrier MA schemes such as OFDMA and SC-FDMA. Blocks of data represented by vector $\mathbf{s}$ of size $M \times 1$ are precoded with an $M \times M$ precoding matrix $\mathbf{P}$. The $M \times 1$ output vector is then mapped to M out of N inputs of the inverse-DFT block according to the subcarrier mapping $N \times M$ transform matrix $\mathbf{\Omega}$. To overcome the effects of frequency-selective channel fading, a CP of length N_{CP} is appended to beginning of each $N \times 1$ block output by the inverse-DFT function. Transmission with different rates among users is available according to each user's requirement, as a different number of

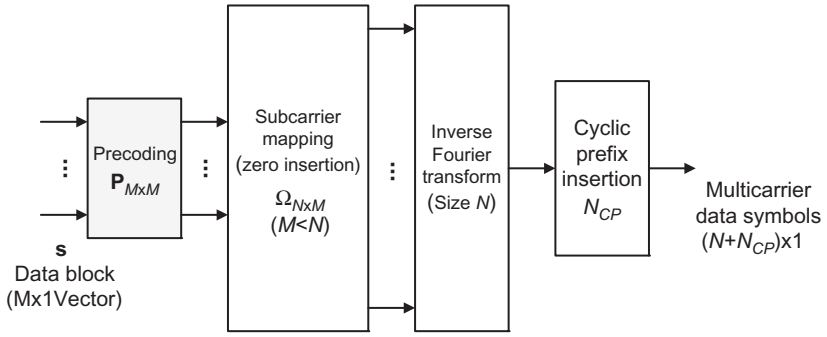


Figure 3.30
General multicarrier transmission scheme [47].

subcarriers and a different modulation and coding schemes can be applied to each user. Let $\mathbf{s}(n)$ denote the information symbols which are parsed into data blocks of size M. The ith data block $\mathbf{s}_i$ can be written as $\mathbf{s}_i = [s(iM), \ldots, s(iM+M-1)]^T$. Let's further denote by $\otimes$ the Kronecker product, by O the all-zero matrix of size $M \times N$ and by $\mathbf{I}$ the $M \times M$ identity matrix. We assume that the size of the inverse DFT is a multiple of the block size $N = MK$.

The special case of $\mathbf{P} = \mathbf{I}$ results in OFDMA where the user-specific data blocks are mapped to a subset of $M < N$ subcarriers, which are selected by the user-specific subcarrier mapping matrix $\mathbf{\Omega}$. The vector $\mathbf{\Omega s}$ is fed to the inverse-DFT function. The form of the matrix $\mathbf{\Omega}$ might lead to either a localized or a distributed subcarrier mapping as follows:

$$\mathbf{\Omega}_{N\times M} = \begin{pmatrix} \mathrm{O}_{L\times M} \\ \mathbf{I}_{M\times M} \\ \mathrm{O}_{(N-L-M)\times M} \end{pmatrix} \quad \text{or} \quad \mathbf{\Omega}_{N\times M} = \mathbf{I}_{M\times M} \otimes \begin{pmatrix} \mathrm{O}_{n\times 1} \\ 1 \\ \mathrm{O}_{(K-n-1)\times 1} \end{pmatrix}$$

By assigning different groups of subcarriers to different users, each user's transmit power can be concentrated in a restricted part of the channel bandwidth, resulting in significant coverage enhancement. Different user signals remain orthogonal only if time/frequency synchronization is maintained and an appropriate CP is appended to compensate for timing misalignments at the receiver. In order to maintain good performance in frequency-selective fading channels, robust forward error correction schemes must be employed.

Precoded OFDMA is a variant of OFDMA in which a precoding matrix $\mathbf{P}$ is used that spreads the energy of symbols over the subcarriers allocated to the user. Uniform energy distribution is desirable in practice. One of the most well-known precoding matrices is the Walsh–Hadamard matrix $\mathbf{P} = \begin{pmatrix} \mathbf{p}_0 & \mathbf{p}_1 & \ldots & \mathbf{p}_{M-1} \end{pmatrix}^T$ where the row vectors $\mathbf{p}_i$ are orthogonal Walsh–Hadamard sequences of length M [47].

3.3.2 Single-Carrier Frequency Division Multiple Access

SC-FDMA has been used as the uplink multiple-access scheme in LTE systems. The use of SC-FDMA was motivated by the fact that a single-carrier system with an OFDMA-type multiple-access would combine the advantages of the two techniques, that is, low PAPR and large coverage. The first SC-FDMA concept was interleaved FDMA (IFDMA), which was based on compression and block repetition of the modulated signal in the time domain. It can be theoretically shown that the spectrum of the compressed and K times repeated signal has the same shape as the original signal, with the difference that it presents exactly $K-1$ zeros between two data subcarriers. This feature enables us to easily interleave different users in the frequency domain by applying to each user a specific frequency shift, or equivalently, by multiplying the time-domain sequence by a specific phase ramp. In addition, similar to OFDMA, robustness to inter-cell interference can be achieved by coordinating resource allocation between adjacent cells. The same waveform can be obtained in the frequency domain, if discrete Fourier transform matrix $\mathbf{P} = [p_{k,n}]; p_{k,n} = \exp(j2\pi kn/M)$ is used as the precoding matrix in Fig. 3.30, resulting in DFT-precoded OFDMA, which is mathematically identical to IFDMA in a distributed scenario. The precoding operation $\mathbf{Ps}$ is equivalent to an $M-$ point DFT operation. With a subcarrier mapping matrix $\mathbf{\Omega}$ as given in Section 3.3.1, the spectrum of the distributed DFT-precoded OFDMA signal is identical to the IFDMA signal spectrum, thus it corresponds to the same waveform. This is also called DFT-spread OFDM. The two techniques are different implementations of SC-FDMA. The advantage of DFT-precoded OFDMA is in its more flexible structure. While IFDMA imposes a distributed signal structure, DFT-precoded OFDMA allows the use of an appropriate subcarrier mapping matrix $\mathbf{\Omega}$. Localized variants of implementation or channel-dependent mappings are also possible. A pulse-shaping filter can be further applied in the frequency domain, with a lower complexity than the time-domain filtering. Note that in frequency-selective channel scenarios, interference may occur among the M elements of each data block. This degradation, which is more important in a distributed subcarrier mapping, also impacts Walsh–Hadamard precoded OFDMA [47]

3.3.3 Non-orthogonal Multiple-Access Schemes

Previous generations of cellular standards relied on orthogonal MA, where each time/frequency resource block was exclusively assigned to one of the active users to ensure no inter-user interference would occur. In 3GPP NR, synchronous/scheduling-based orthogonal MA continues to play an important role in uplink/downlink transmissions. Non-orthogonal multiple access transmission, which allows multiple users to share the same time/frequency resource, was recently proposed to enhance the system capacity and to accommodate massive connectivity through asynchronous uplink access. Unlike orthogonal MA, multiple

non-orthogonal multiple access users' signals are multiplexed using different power allocation coefficients or different signatures such as codebooks/codewords, sequences, interleavers, and preambles. The fundamental theory of non-orthogonal multiple access has been extensively studied in network information theory. The uplink and downlink non-orthogonal multiple access can be theoretically modeled as a multiple-access channel and a broadcast channel, respectively. The capacity region of the Gaussian broadcast channel can be achieved by power-domain superstition coding with a successive interference cancellation (SIC) receiver. Meanwhile, the capacity region of Gaussian multiple-access channel corresponds to CDMA, where different codes are used for the different transmitters, and the receiver decodes them in an SIC manner. In general, a user with poor-channel condition tends to allocate more transmission power, so this user would decode its own messages by treating the co-scheduled user's signal as noise. On the other hand, a user with good channel condition applies the SIC strategy by first decoding the information of the poor-channel user and then decoding its own, removing the other users' information. The results of studies in 3GPP suggest that using a non-SIC receiver results in negligible performance degradation in many cases [23]. Relaxing the need for an SIC receiver significantly would reduce the decoding complexity for the downlink case as the others' codebooks are no longer required.

In addition to the orthogonal MA scheme, the 3GPP NR may support an uplink non-orthogonal transmission (*Note: at the time of publication of this book, 3GPP has decided not to specify non-orthogonal multiple access in Rel-16 and instead only specify a two-step RACH*) to provide the massive connectivity that is desperately required for applications in mMTC as well as other scenarios. 3GPP has further studied grant-free uplink multiple-access schemes for mMTC scenarios. Since there is no need for a dynamic and explicit scheduling grant from gNB, latency reduction and control signaling minimization could be expected. For uplink non-orthogonal multiple access, network information theory suggests that CDMA with a SIC receiver provides a capacity achieving scheme. However, securing uplink non-orthogonal multiple access gain requires further system design enhancement. As the number of co-scheduled users increases, the decoding complexity of the SIC receiver increases. The message passing algorithm (MPA), as a less-complex decoding algorithm, as well as other low-complexity receiver designs have recently drawn attention. Several code-spreading-based techniques, including SCMA, MUSA, and PDMA and several others were the candidates under consideration in 3GPP Rel-16 NOMA study item. It has been shown that one can potentially achieve higher spectral efficiency, larger connectivity, and better user fairness with non-orthogonal multiple access relative to orthogonal MA schemes [50–52].

While the interference-free condition between orthogonally multiplexed users might facilitate multi-user detection at the receivers, it is widely known that orthogonal MA cannot achieve the sum capacity of a wireless system. The orthogonal MA also has limited granularity of resource scheduling, so it struggles to handle a large number of active connections. Non-orthogonal multi-user transmission/access has been recently investigated in a

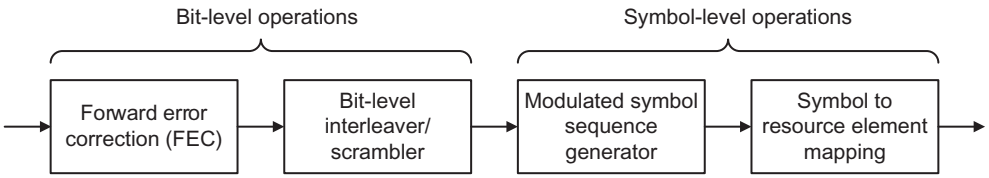


Figure 3.31
High-level block diagram for uplink non-orthogonal multiple-access schemes [22,23].

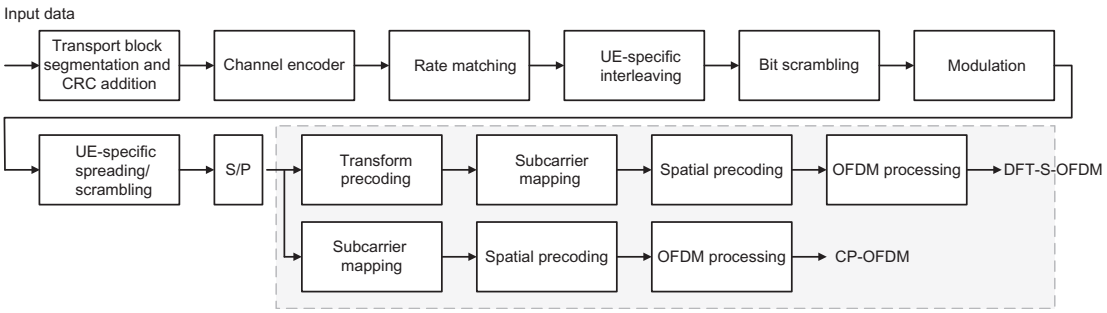


Figure 3.32
General framework for non-orthogonal multiple access uplink transmission [23].

systematic manner to deal with the above problems. Interference can be controlled by non-orthogonal resource allocation at the cost of increased receiver complexity.

The non-orthogonal multiple access schemes that were studied for the uplink transmission in 3GPP have the following features in common: (1) they use MA signature(s) at the transmitter side and (2) they allow multi-user detection at the receiver side. The MA signatures are typically used to differentiate the users. Thus all proposed non-orthogonal MA schemes for the uplink transmission at a high level can be described with the basic diagram shown in Fig. 3.31.

As we stated earlier, in non-orthogonal multiple access uplink transmission, multiple UEs share the same time/frequency resources via non-orthogonal resource allocation. There are various non-orthogonal multiple access schemes that can be derived from the general concept shown in Fig. 3.31. Fig. 3.32 shows a unified framework for non-orthogonal multiple access based on UE-specific spreading/scrambling/interleaving for the uplink data transmission. For data transmission based on UE-specific spreading, the existing solutions can be classified into two categories: linear spreading and nonlinear spreading. The category of linear spreading includes solutions such as resource spread multiple access (RSMA), MUSA, WSMA, and NCMA, while the category of nonlinear spreading includes SCMA. Linear spreading can be used in conjunction with DFT-S-OFDM and CP-OFDM waveforms. Although receiver implementation relatively less-complex and straightforward in orthogonal MA systems, the successful deployment of non-orthogonal multiple access depends on advanced receivers with inter-UE interference cancellation capabilities [23].

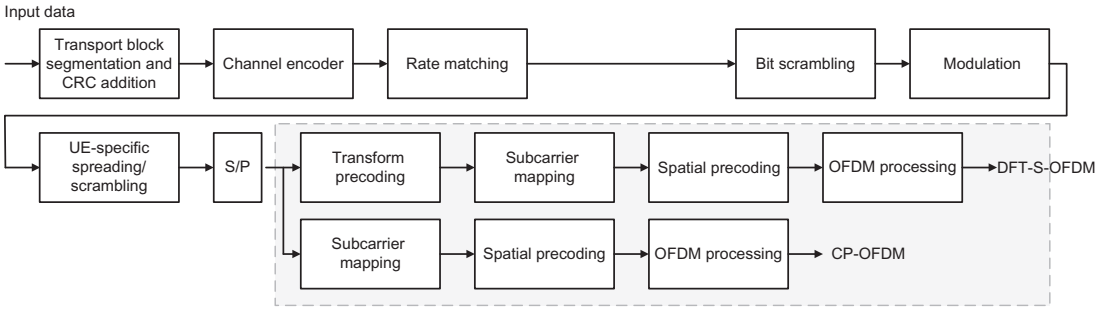


Figure 3.33
Linear hybrid spreading for non-orthogonal multiple access uplink transmission [23].

Fig. 3.33 shows non-orthogonal multiple access uplink transmission blocks based on linear hybrid spreading. In particular, the assignment of linear spreading codes is UE specific, which carries the MA signature. The assignment of scrambling sequence can be UE or cell specific. Same or different set of spreading codes and scrambling sequences can be employed for CP-OFDM and DFT-S-OFDM waveforms. To randomize the inter-UE interference and maximize the reuse of non-orthogonal multiple access resources, the mapping of spreading code and scrambling sequence can be made symbol dependent.

Compared to nonlinear spreading schemes such as SCMA, the studies indicate that solutions based on linear hybrid spreading exhibit similar BLER and significantly better performance in terms of scalability, complexity, PAPR, and flexibility. The use of long spreading codes can obtain a large codebook with good autocorrelation and low cross-correlation properties, which is suitable for grant-free non-orthogonal multiple access and is further robust against transceiver synchronization errors and timing inaccuracies associated with asynchronous transmission. The large processing gain of long spreading codes is also beneficial for inter-UE interference suppression. Therefore, it is an obvious candidate for grant-free, asynchronous transmission, relaxing the timing advance requirements. To improve the spectral efficiency, multi-layer transmission can be used, either in the form of standalone or in combination with spatial multiplexing schemes (in the presence of multiple transmit antennas). Compared to long spreading codes, short spreading codes need smaller spreading factor and higher spectral efficiency. The short spreading codes can be optimized to achieve the Welch bound[15] on cross-correlation, which can be leveraged for multi-user detection and inter-UE

[15] In mathematics, Welch bounds are a family of inequalities corresponding to the problem of evenly spreading a set of unit vectors in a vector space. If $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ are unit vectors in $\mathbb{C}^n$. We define $c_{\max} = \max_{i \neq j} \left| \langle \mathbf{x}_i, \mathbf{x}_j \rangle \right|$ where the inner product is defined in $\mathbb{C}^n$. Then, the following inequalities are given $(c_{\max})^{2k} \geq 1/(m-1)\left[m/\left(\binom{n+k-1}{k} - 1 \right) \right], \forall k = 1, 2, \ldots$. If $m \leq n$, then the vectors $\{\mathbf{x}_n\}$ can form an orthonormal set in $\mathbb{C}^n$. In this case, $c_{\max} = 0$ and the bounds are void. Consequently, interpretation of the bounds is only meaningful if $m > n$.

interference cancellation for synchronized reception. Moreover, it can be easily combined with spatial precoding to further mitigate the cross-correlation and enhance the non-orthogonal multiple access capacity.

A large number of the existing non-orthogonal multiple access schemes are based on linear spreading. For this type of spreading-based NOMA schemes, the same type of NOMA receivers can be utilized. The NOMA receiver consists of three parts. The first part is multi-user detector where the superposed received signal is processed jointly across the UEs to derive the LLR for each UE. The second part is the channel decoder which receives the LLR from multi-user detector and decodes the transmitted codeword. The output from the channel decoder can be either a decoded codeword when the decoding is successful or an intermediate LLR for each bit refined through the message passing decoding process. The third part is the iteration between the multi-user detector and the channel decoder, where they can exchange both soft-LLR information and hard decision information. We will focus on multi-user detector here and in particular, we will focus on elementary signal estimator (ESE) and linear MMSE (LMMSE) estimator. Without loss of generality, we consider single symbol processing. Let's assume that there are J users and K resources (spreading factor). The received signal at resource k can be written as $y_k = \sum_{j=1}^{J} h_{kj}x_{kj} + n_k$ where h_{kj} is the channel coefficient corresponding to resource k from user j, x_{kj} is the transmitted signal by user j on resource k, and $n \sim \mathbb{C}(0, N_0)$. For linear spreading codes, each user is assigned a spreading code sequence. Let c_{kj} be the kth coefficient of the spreading code for user j. We assume that all users share the same modulation alphabet $W = \{w_1, w_2, \ldots, w_M\}$. Then, $x_{kj} = c_{kj}s_j \forall k = 1, 2, \ldots, K$ where $s_j \in W$ is the transmitted symbol by user j. The received signal can be written as $\mathbf{y = Hs + n}$ where $\mathbf{s} = [s_1 \ldots s_J]^T$, $\mathbf{n} = [n_1 \ldots n_J]^T$, and $\mathbf{H} = [\tilde{h}_{kj}]_{k,j}$ is a $K \times J$ matrix with entries $\tilde{h}_{kj} = h_{kj}c_{kj}$. Let $\tilde{h}_j$ denote the jth column of matrix $\mathbf{H}$. The multi-user detector estimates the LLRs for $\mathbf{s}$ based on $\mathbf{y}$ [23].

Matched filter and ESE, which is a generalization of the matched filter that can accommodate soft interference cancellation, can be used as multi-user detectors. The ESE multi-user detector first compresses the received signals to scalar values for each UE by matched filtering. The output of the matched filter can be written as $\hat{\mathbf{y}} = [\hat{y}_1 \ldots \hat{y}_J]^T = \mathbf{H}^H\mathbf{y}$. In order to take advantage of soft information computed at the channel decoder, we can apply the ESE to $\hat{\mathbf{y}}$, which approximates signal and interference as Gaussian random variables as follows $\hat{y}_j = \left|\hat{\mathbf{h}}_j\right|^2 s_j + \sum_{i \neq j} \hat{\mathbf{h}}_j^H \hat{\mathbf{h}}_i s_i + \hat{\mathbf{h}}_j^H \mathbf{n} = \left|\hat{\mathbf{h}}_j\right|^2 s_j + \varepsilon_j$ where ε_j is residual interference plus noise. The residual interference ε_j is approximated as a Gaussian random variable which can be described by its mean and variance as follows:

$$\mu(\varepsilon_j) = \sum_{i \neq j} \hat{\mathbf{h}}_j^H \hat{\mathbf{h}}_i \mu(s_i), \qquad \sigma^2(\varepsilon_j) = \sum_{i \neq j} \left|\hat{\mathbf{h}}_j^H \hat{\mathbf{h}}_i\right|^2 \sigma^2(s_i) + \left|\hat{\mathbf{h}}_j\right|^2 N_0$$

where $\mu(s_i)$ and $\sigma^2(s_i)$ are the a priori mean and variance of the symbol transmitted from the ith user, which can be computed using a priori bit LLRs. The LMMSE estimation can be used to estimate s_j from $\hat{y}_j$. From the estimation, LLR for each bit can be derived from conventional marginalization.[16] It can be noted that the ESE multi-user detector without matched filter and symbol spreading can also be used. For random symbol interleaver cases, the ESE multi-user detector can be applicable assuming that $p(\mathbf{y}|s_i = s) = \prod_{k=1}^{K} p(y_k|s_i = s)$. The ESE multi-user detector is also applicable to bit-level interleaving scenarios. It can be shown that the computational complexity of ESE multi-user detector scales as $O(K^2)$ for J UEs and spreading factor K [23].

Unlike ESE multi-user detector, the LMMSE estimator treats the received signal as a vector and applies LMMSE estimation matrix to the transmitted signal estimation of each UE. Let $\boldsymbol{\mu}_p$and$\mathbf{v}_p$ vectors denote the priori mean and variance of each UE transmitted signal derived from LLRs. The receiver applies the following LMMSE filter to the received vector where the output of the LMMSE filter is the mean and variance vectors.

$$\boldsymbol{\mu} = [\ \mu_1 \quad \ldots \quad m_J\ ]^T = \boldsymbol{\mu}_p + \mathbf{v}_p\mathbf{H}^H(N_0\mathbf{I} + \mathbf{H}\mathbf{v}_p\mathbf{H}^H)^{-1}(\mathbf{y} - \mathbf{H}\boldsymbol{\mu}_p)$$
$$\mathbf{v} = [\ v_1 \quad \ldots \quad v_J\ ]^2$$
$$v_i = v_{ip} - v_{ip}^2\hat{\mathbf{h}}_j^H(N_0\mathbf{I} + \mathbf{H}\mathbf{v}_p\mathbf{H}^H)^{-1}\hat{\mathbf{h}}_j$$

As mentioned earlier, $\boldsymbol{\mu}_p$ is the a priori mean vector and $\mathbf{v}_p$ is a diagonal matrix whose diagonal entries are a priori variance values for the corresponding transmitted symbols. A priori mean and variance values can be computed using the bit LLRs computed at the channel decoder. Based on the LMMSE output, the receiver generates extrinsic bit LLR values for channel decoder by marginalization. It can be shown that the computational complexity of LMMSE multi-user detector scales as $O(K^3 + K^2J + KJ^2)$ for J UEs and spreading factor of K [23].

Table 3.6 summarizes the use cases and operation modes of Rel-16 NOMA candidates. The features in the characteristics column reflect the potential benefits of NOMA over Rel-15 NR MA scheme [6], which are considered in the design, evaluation and comparison of NOMA transmitter and receiver schemes. The studies conducted in 3GPP suggest that when

[16] In probability theory and statistics, the marginal distribution of a subset of a collection of random variables is the probability distribution of the variables contained in the subset. It gives the probabilities of various values of the variables in the subset without reference to the values of the other variables. This contrasts with a conditional distribution, which gives the probabilities contingent upon the values of the other variables. Marginal variables are those variables in the subset of variables being retained. The distribution of the marginal variables (the marginal distribution) is obtained by marginalizing; that is, focusing on the sums in the margin, over the distribution of the variables being discarded, and the discarded variables are said to have been marginalized.

Table 3.6: Rel-16 non-orthogonal multiple-access use cases and features supported by different operation modes [23].

Operation Mode		Dynamic MCS Support	Characteristics	Use Case
RRC_INACTIVE, grant-free with contention, tracking area-free (asynchronized)		No	Reduction in system overhead, latency, and power consumption	mMTC and eMBB
RRC_CONNECTED (Synchronized)	Grant-free with contention	No	Reduction in system overhead and latency	mMTC, URLLC, and eMBB
	Grant-based with overloading	Yes	Limited downlink overhead reduction	eMBB

the network operates in grant-based mode, transmission schemes proposed for NOMA can be applied to MU-MIMO; however, the relative spectral efficiency advantage of NOMA over MU-MIMO is not clear under underloaded scenarios. When the network operates in grant-free mode and the uplink access is contention-free, the relative gain of NOMA over MU-MIMO in terms of spectral efficiency is not proven. The most significant gain of NOMA over MU-MIMO may be attributed to contention-based, grant-free transmission and small data transmission in RRC_INACTIVE state scenarios.

3.3.3.1 Sparse Code Multiple Access

SCMA is a frequency-domain non-orthogonal multiple-access scheme, which can improve the spectral efficiency of wireless radio access. In SCMA, different incoming data streams are directly mapped to codewords of different multi-dimensional cookbooks, where each codeword represents a spread transmission layer. Each layer or user has its own dedicated codebook. Multiple SCMA layers share the same time-frequency resources of OFDMA. The sparsity of codewords makes the near-optimal detection feasible through iterative MPA.[17] Such low complexity of multi-layer detection allows excessive codeword overloading in which the dimension of multiplexed layers exceeds the dimension of codewords.

[17] Belief propagation, also known as sum-product message passing, is a message-passing algorithm for performing inference on graphical models such as Bayesian networks and Markov random fields. It calculates the marginal distribution for each unobserved node conditional on any observed nodes. Belief propagation is commonly used in artificial intelligence and information theory and has demonstrated empirical success in numerous applications including low-density parity-check codes, turbo codes, free energy approximation, and satisfiability. The algorithm was formulated as an exact inference algorithm on trees, which was later extended to polytrees. While it is not accurate for general graphs, it has been shown to be a useful approximation algorithm. If $X = \{X_i\}$ is a set of discrete random variables with a joint mass function p, the marginal distribution of a single X_i is the summation of p over all other variables $p_{X_i}(x_i) = \sum_{\mathbf{x}': x'_i \neq x_i} p(\mathbf{x}')$. However, this would become computationally prohibitive, whereas by exploiting the polytree structure, belief propagation would allow the marginals to be computed more efficiently [50–52].

Optimization of overloading factor along with modulation/coding levels of layers provides a more flexible and efficient link-adaptation mechanism. On the other hand, the signal spreading feature of SCMA can improve link-adaptation as a result of less colored interference. In SCMA, incoming bits are directly mapped to multi-dimensional complex codewords selected from predefined codebook sets. The co-transmitted spread data are carried over super-imposed layers. Since layers are not fully separated in a NOMA system, a nonlinear receiver is required to detect the intended layer of each user. Therefore, additional detection complexity is the cost of the nonorthogonal multiple-access especially when the system is heavily overloaded with a large number of multiplexed layers. Low-density spreading (LDS) is a special form of SCMA. In LDS, codewords are built by spreading of modulated symbols using LDS signatures with few non-zero elements within a large signature length. Despite the moderate complexity of detection, LDS suffers from poor performance especially for large constellation sizes beyond QPSK. All CDMA schemes, and in particular LDS, can be considered as different types of repetition coding in which different variations of a QAM symbol are generated by a spreading signature. Repetition coding is not able to provide desirable spectral efficiency for a wide range of SNR. In SCMA the QAM mapper and linear operation of sparse spreading are merged to directly map incoming bits to a complex sparse vector called a codeword. Both LDS and SCMA are based on the idea that one-user information is spread over multiple subcarriers. However, the number of subcarriers assigned to each user is smaller than the total number of subcarriers, and this low spreading (sparse) feature ensures that the number of users utilizing the same subcarrier is not too large, such that the system complexity remains manageable [50–52].

In LDS, each user spreads its data on a small set of subcarriers. There is no exclusivity in the subcarrier allocation and more than one user can share each subcarrier. The interference pattern at the receiver will generate a low-density graph, and graph theory-based techniques can be utilized. The main features of the LDS scheme can be summarized as follows. At each subcarrier, a user will have relativity small number of interferers comparing to the total number of users. Consequently, the search space will be smaller and more complex multi-user detection techniques can be implemented. Higher SINR can be achieved at each subcarrier, which results in reliable detection process. Each user will experience interference from different users at different subcarriers, which results in interference diversity by avoiding strong interferers to destroy the signal of a user on all the subcarriers. Belief propagation based multi-user detection can be implemented with linear complexity in the number of subcarriers [50–52].

The LDS and SCMA share the same concept which is to use a low-density or sparse non-zero element sequence to reduce the complexity of MPA processing at the receiver. However, in SCMA, bit streams are directly mapped to different sparse codewords. An example is illustrated in Fig. 3.34, where each user has a codebook and there are six users. All codewords in the same codebook contain zeros in the same two dimensions, and the positions

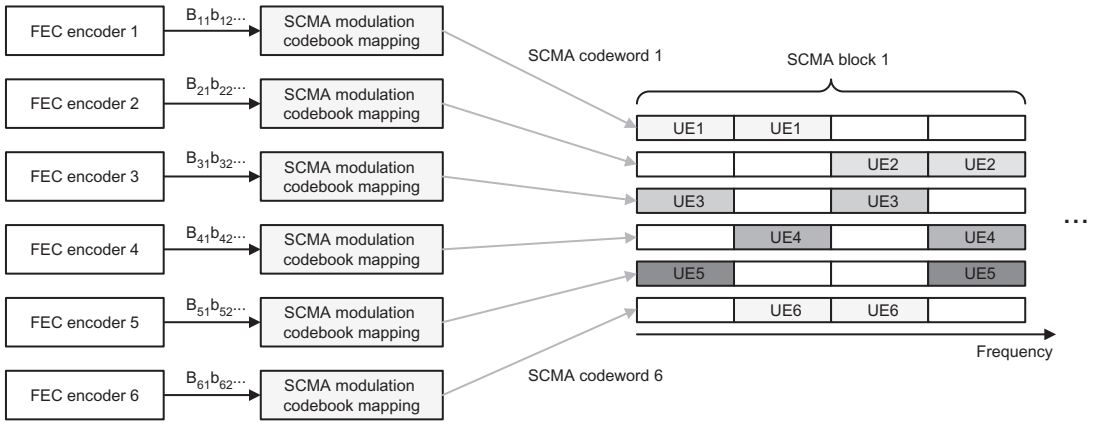


Figure 3.34
Illustration of SCMA concept [52].

of the zeros in the different codebooks are distinct to facilitate the collision avoidance of any two users. For each user, two bits are mapped to a complex codeword. Codewords for all users are multiplexed over four shared orthogonal resources. The key difference between LDS and SCMA is that a multi-dimensional constellation for SCMA is designed to generate codebooks, which provides a shaping gain that is not possible with LDS. In order to simplify the design of the multi-dimensional constellation, a baseline constellation can be generated by minimizing the average alphabet energy for a given minimum Euclidian distance between constellation points, and also taking into account the codebook-specific operations such as phase rotation, complex conjugate, and dimensional permutation [50–52].

We use the uplink multiple-access system shown in Fig. 3.35 to explain the SCMA processing stages. Let's assume that the system comprises K users whose information bits are spread over N resource elements. In an orthogonal scenario, $K \leq N$ to ensure that each user is assigned to an orthogonal resource element, while in the non-orthogonal scenarios, $K > N$ and the ratio K/N is defined as the overloading factor. The transceiver structure of SCMA can be mathematically modeled as follows. Let $\mathbf{b} = \begin{bmatrix} b_1 & b_2 & \dots & b_K \end{bmatrix}^T$ denote the information bits transmitted by K uplink users and $\mathbf{x}_k = [x_k^1 \; x_k^2 \; \dots \; x_k^N]^T$ represent the transmitted symbols by the kth user. An SCMA encoder at the kth user is defined to be a one-to-one mapping $f_k : B_k \rightarrow X_k$ with $b_k \in B_k$ and $\mathbf{x}_k \in X_k$, where the cardinality of B_k and X_k is 2^{N_B} where N_B denotes the number of information bits in b_k. Note that due to the sparse nature of SCMA scheme, $\mathbf{x}_k$ may contain zero symbols. The received signal at the base station $\mathbf{y}$, after passing through a block fading multiple-access (uplink) channel, can be expressed as $\mathbf{y} = \sum_{k=1}^{K} \mathbf{H}_k \mathbf{x}_k + \mathbf{z}$ where $\mathbf{H}_k = \text{diag}(\mathbf{h}_k^1 \quad \mathbf{h}_k^2 \quad \dots \quad \mathbf{h}_k^N)$ represents the channel between the base station and the kth user, $\mathbf{z} = \begin{bmatrix} z_1 & z_2 & \dots & z_N \end{bmatrix}$ is the additive white Gaussian

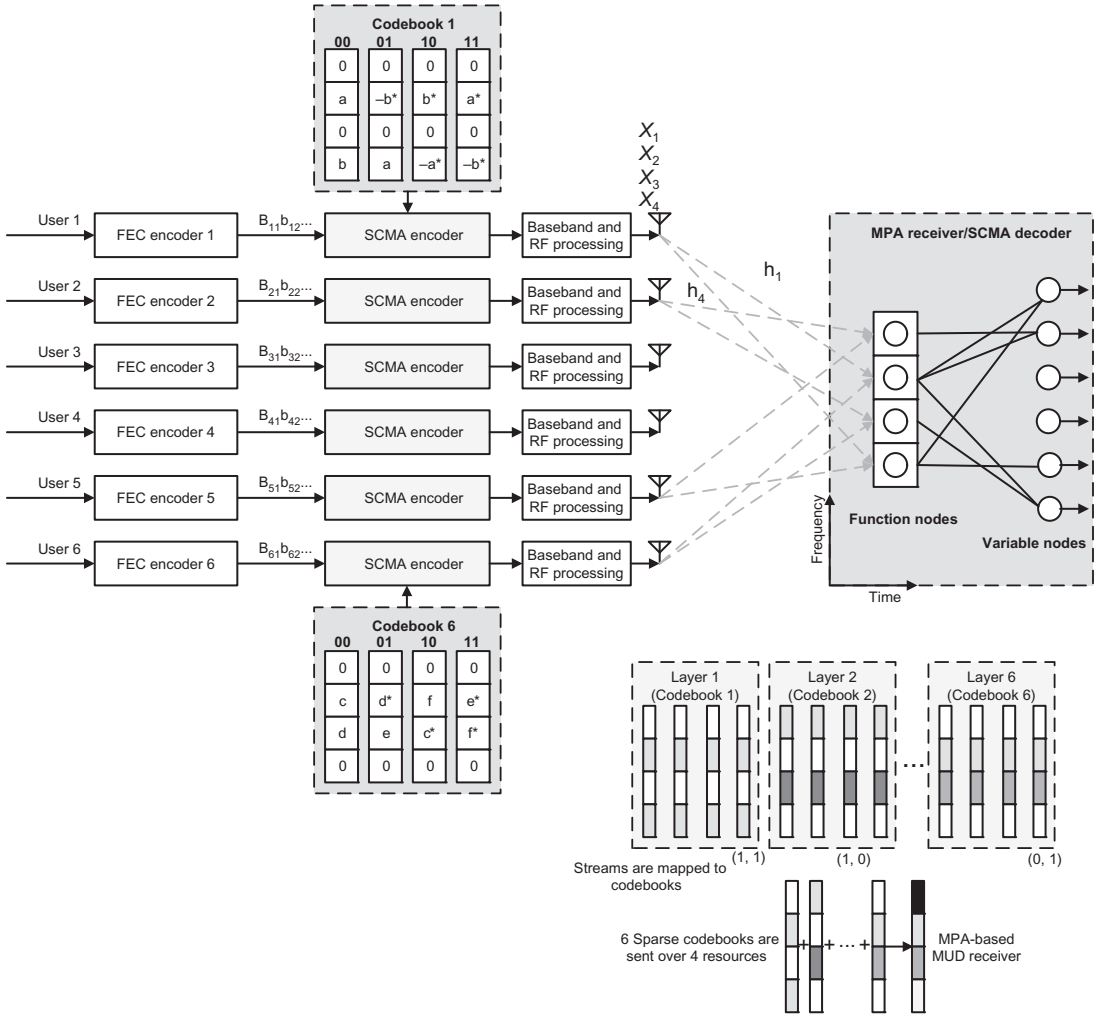


Figure 3.35
Example SCMA processing with $K = 6, N = 4$ and 150% overloading factor [52].

noise with zero mean and unity variance. Given the received signal $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \dots & y_N \end{bmatrix}$ and the channel knowledge $\mathbf{H} = \{\mathbf{H}_k | k = 1, 2, \dots, K\}$, the joint maximum-A-posteriori detection[18] of $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_K \end{bmatrix}$ can be written as $\hat{\mathbf{X}} = \arg \max_{\mathbf{X} \in X_1 \times X_2 \times \cdots \times X_K} p(\mathbf{X}|\mathbf{y})$. In general, the solution to the above problem requires a global search over the joint space of K

[18] Maximum-a-posteriori (MAP) estimate of random variable X given that we have observed $Y = y$, is given by the value of x that maximizes $f_{X|Y}(x|y)$ if X is a continuous random variable, and $p_{X|Y}(x|y)$ if X is a discrete random variable. The MAP estimate is shown by $\hat{x}_{MAP} = \arg\max_x f_{X|Y}(x|y)$.

uplink users $X_1 \times X_2 \times \cdots \times X_K$. Due to the sparse nature of SCMA transmission scheme, the MPA detector can be applied to reduce the decoding complexity, which iteratively updates the belief associated with the underlying factor graph. Once $\hat{\mathbf{X}}$ has been estimated, we can use the inverse mapping function $(f_k)^{-1}$ to recover the original user information bits B_k [50–52].

3.3.3.2 Power-Domain Non-orthogonal Multiple Access

Power-domain NOMA can serve multiple users in the same time slot, OFDMA subcarrier, or spreading code, and multiple-access is realized by allocating different power levels to different users depending on their relative position to the base station.

Fig. 3.36 illustrates the concept of power-domain NOMA in the downlink with two UEs that utilize SIC receiver. For simplicity, we assume in this section the case of single transmit/receive antennas. The overall system transmission bandwidth is assumed to be 1 Hz. The base station transmits signal x_i to the ith UE with transmit power P_i where $E\{|x_i|^2\} = 1$ assuming $\sum_i P_i = P$. In power-domain NOMA, x_1 and x_2 are superposed as $x = \sqrt{P_1}x_1 + \sqrt{P_2}x_2$; thus the received signal at the ith UE can be written as $y_i = h_i x + w_i$, in which h_i is the complex-valued channel coefficient between the ith UE and the base station, and w_i denotes a zero-mean AWGN plus inter-cell interference. The PSD of w_i is N_{0i}. In the downlink NOMA, the SIC receiver is implemented at the UE receiver. The optimal order for decoding is in the order of decreasing channel gain normalized by noise and

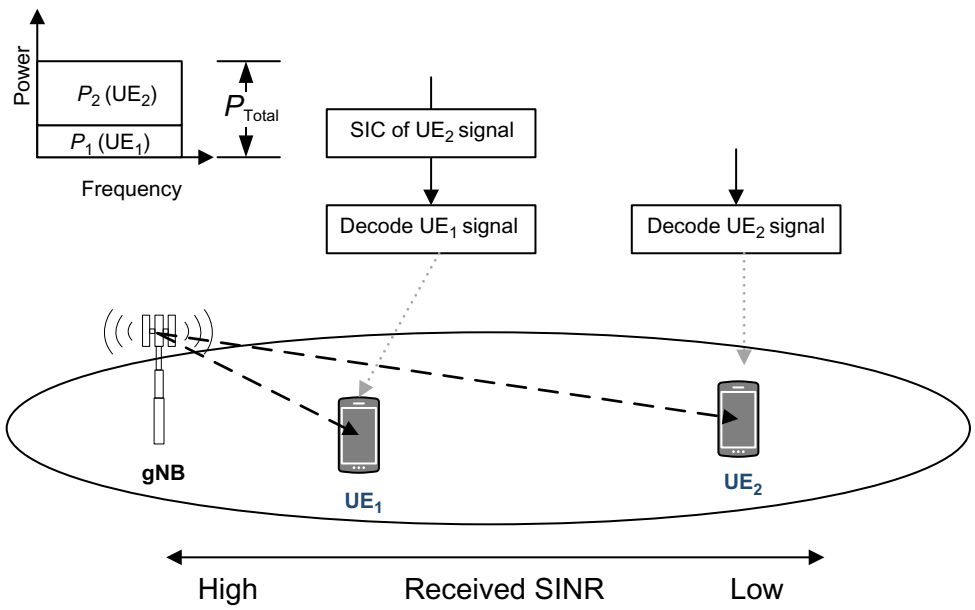


Figure 3.36
Illustration of the principle of downlink power-domain NOMA [52].

inter-cell interference power, that is, $|h_i|^2/N_{0i}$. Based on this order, we assume that any user can correctly decode the signals of other users whose decoding order comes before the corresponding user. Therefore, the ith UE can remove the inter-user interference from the jth user whose channel gain $|h_j|^2/N_{0j}$ is less than $|h_i|^2/N_{0i}$. In the example with two UEs, assuming that $|h_1|^2/N_{01} > |h_2|^2/N_{02}$, UE_2 does not have to perform interference cancellation since it comes first in the decoding order. UE_1, on the other hand, has to first decode x_2 and subtract that component from received signal y_1, then to decode x_1 without interference from x_2. Assuming successful decoding and no error propagation, the throughput of the ith UE, R_i, is given as follows [52]:

$$R_1 = \log_2\left(1 + \frac{P_1|h_1|^2}{N_{01}}\right) \quad \text{and} \quad R_2 = \log_2\left(1 + \frac{P_2|h_2|^2}{P_1|h_2|^2 + N_{02}}\right)$$

It can be seen that power allocation for each UE greatly affects the user throughput and thereby the modulation and coding scheme used for data transmission of each UE. By adjusting the power allocation ratio P_1/P_2, the base station can effectively control the throughput of each UE. The overall cell throughput, cell-edge throughput, and user fairness are closely related to the adopted power allocation scheme [52].

In a system that uses orthogonal MA scheme and hypothetically serves two UEs, if normalized bandwidth $0 < \beta < 1$ is assigned to the first UE, the remaining bandwidth $1 - \beta$ will be assigned to the second UE to maintain orthogonality between the users. The throughput of the ith UE R_i is given as follows:

$$R_1 = \beta\log_2\left(1 + \frac{P_1|h_1|^2}{\beta N_{01}}\right) \quad \text{and} \quad R_2 = (1-\beta)\log_2\left(1 + \frac{P_2|h_2|^2}{(1-\beta)N_{02}}\right)$$

In power-domain NOMA, the performance gain relative to orthogonal MA increases when the difference in channel gains, for example, path loss between UEs, is large.

The uplink capacity can be calculated in the similar manner as the downlink, although the formula is somewhat different. Defining P_{r1} and P_{r2} as the received powers at the base station from UE_1 and UE_2, respectively, the rate of each user in the case of non-orthogonal uplink access can be written as follows:

$$R_1 < \log_2\left(1 + \frac{P_{r1}}{N_0}\right) \quad \text{and} \quad R_2 < \log_2\left(1 + \frac{P_{r2}}{N_0}\right)$$
$$R_1 + R_2 < \log_2\left(1 + \frac{P_{r1} + P_{r2}}{N_0}\right)$$

Multicarrier NOMA can be viewed as a variation of NOMA, where the users in a network are divided into multiple groups. The users in each group are served in the same orthogonal resource block following the NOMA principle, and different groups are allocated to

different orthogonal resource blocks. The motivation for employing hybrid NOMA is to reduce the system complexity. For example, assigning all the users in the network to a single group for the implementation of NOMA in one orthogonal resource block is problematic, since the user having the best channel conditions will have to decode all the other users' messages before decoding its own message, which results in high-complexity and high-decoding delay. Hybrid NOMA is an effective approach to make a tradeoff between system performance and complexity. Let's consider multicarrier NOMA as an example. The users in the cell are divided into multiple groups which are not necessarily mutually exclusive. The users within one group are assigned to the same subcarrier, and intra-group interference is mitigated using the NOMA principle. Different groups of users are assigned to different subcarriers, which effectively avoids inter-group interference. As a result, overloading the system, which is necessary in order to support more users than the number of available subcarriers and is required to enable massive connectivity, can be realized by the hybrid NOMA scheme. It is noted that, with hybrid NOMA, overloading is realized at reduced complexity since the number of users assigned to each subcarrier is limited [50–52].

The base station scheduler in power-domain NOMA searches and pairs multiple users for simultaneous transmission at each subband. To determine the set of paired users and the allocated power set at each subband, a multi-user proportional fairness (PF) scheduler may be used. The PF scheduling metric attempts to find the candidate user sets U and power sets P_s that maximize the following expression over each subband s:

$$Q(U, P_s) = \sum_{k \in U, P_s} \left(\frac{\eta(k, U, P_s, t)}{L(k, t)} \right)$$

where $(U_{\max}, P_{s_{\max}}) = \max_{U, P_s} Q(U, P_s)$ denotes the maximum argument of PF scheduling metric $Q(U, P_s)$ for candidate user set U and allocated power set P_s over all users in the user set, $\eta(k, U, P_s, t)$ is the instantaneous throughput of user k in subband s at time instance t (the time index of a subframe), whereas $L(k, t)$ is the average throughput of user k. For power-domain NOMA, if we assume the possibility of dynamic switching between NOMA and orthogonal MA, then NOMA can be used only when it provides performance gains. Moreover, the number of users to be multiplexed over each subband is decided by searching all possible candidate user sets of different sizes up to m. The number of candidate user sets to be searched is given by [50–52]:

$$u = \binom{K}{1} + \binom{K}{2} + \cdots + \binom{K}{m}$$

In orthogonal MA schemes, the same MCS is selected over all subbands allocated to a single user. Therefore, the average signal-to-interference plus noise power ratio over all

allocated subbands is used for MCS selection. However, when power-domain NOMA is utilized over each subband, user pairing and power allocation are performed over each subband. With such a mismatch between wideband MCS selection and subband power allocation granularities, the full-scale NOMA gains would not be realized. Furthermore, the higher the power allocation granularity, the more signaling overhead and thus performance degradation.

Power control of uplink NOMA is different from that of downlink in two aspects. In the downlink direction, the transmission power is limited by the maximum transmission power of the base station; however, in the uplink, the transmission power optimization is constrained by the maximum transmission power of individual UEs. In addition, there is a different approach to transmit power control in the uplink. In the downlink, the superposed signal received at a UE experiences the same channel, that is, the signals of different UEs have the same channel gain at each UE receiver. Therefore, the design of downlink power control tries to create sufficiently large difference among the signals of different UEs in the power domain in order to enable signal separation at the (SIC) receiver. For the uplink, due to different channels experienced by signals transmitted via different UEs, the received signal powers of different UEs already have differences in the power domain. On the other hand, when NOMA is applied in the uplink, the ICI greatly increases since multiple UEs are allowed to simultaneously transmit, whereas in the downlink, ICI does not increase when NOMA is applied because generally the base station has fixed transmission power regardless of the number of multiplexed UEs [23].

3.3.3.3 Scrambling-Based and Spreading-Based NOMA Schemes

In addition to the NOMA schemes discussed in the previous sections, there are other schemes proposed as part of 3GPP Rel-16 study item on NOMA. Scrambling-based NOMA schemes use different scrambling signatures for each user and utilize a low-rate forward error correction code or code repetition for multi-user decoding. The scrambling operation is carried out after the modulation. MMSE with SIC (MMSE-SIC) and ESE are used for the multi-user detection. RSMA is one of the scrambling-based schemes under consideration which utilizes low cross-correlation properties of long pseudo-random scrambling codes. Long scrambling sequences are used in RSMA. However, a long user signature causes high-decoding complexity and latency. Following the descrambling step, the ratio of signal-to-interference power is directly proportional to the scrambling code length. It must be noted that each user can transmit signal at any time using the asynchronous RSMA. Depending on the application scenarios, single-carrier RSMA or multicarrier RSMA can be utilized. Single-carrier RSMA can be used in the uplink access to reduce the PAPR of the UE. Multicarrier RSMA can be utilized in the downlink access to simplify the receiver complexity in the frequency-selective wireless fading channels. RSMA can be extended to multiple layers. Treating layers as virtual users, the data is split into multiple parallel layers

for each user. The complexity of multi-layer RSMA is higher than that of single-layer RSMA. The RSMA uses hybrid short-code spreading and long-code scrambling as the MA signatures. The generation of scrambling sequences can be UE-group and/or cell specific, wherein the sequence ID of scrambling code is a function of the cell ID and UE-group ID. One or multiple UE groups can be configured in a cell. The sequences used for scrambling code can be Gold sequences, Zadoff−Chu sequences, or a combination of the two [12,23]. In Welch bound equality (WBE)-based spreading schemes such as RSMA, the design metric for the signature vectors is the total squared cross-correlation $\xi_c \triangleq \sum_{i,j} \left|\mathbf{s}_i{}^H \mathbf{s}_i\right|^2$. The lower bound on the total squared cross-correlation of any set of K vectors of length N is $K^2/N \leq \xi_c$. The WBE sequences are designed to meet the bound on the total squared cross-correlations of the vector set with equality $\xi_{Welch} \triangleq K^2/N$ [23].

The spreading-based NOMA schemes use non-orthogonal short spreading sequences with relatively low cross-correlation, for distinguishing multiple users, and the spreading sequences are non-sparse. The spreading sequences and the decoding algorithm are different for this category of NOMA schemes. In MUSA, modulation symbols of multiple users are spread by specially designed short sequences. All spreading symbols are transmitted over the same time-frequency resources. Multiple spreading sequences constitute a pool from which each user can randomly select. The spreading sequences of MUSA are complex-valued, in which the real part and imaginary part are both drawn from a real-valued multi-level set with uniform distribution. At the receiver, the codeword-level SIC detection is used to separate the target UE signal from the overlapped signals [23].

The average mutual information[19] can be used as a performance metric to compare the spectral efficiencies of various NOMA schemes with OFDMA. This performance metric provides the maximum information rate that can be reliably transmitted for a given channel state information. In a single-user case, the average mutual information is calculated for the signal after constellation mapping and before the soft demapping. This analysis can be extended to the multiuser case for evaluating the achievable sum-rate of the NOMA schemes. Let $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \dots & x_J \end{bmatrix}$ denote the multi-user modulation symbol vector before the NOMA signature pattern, and $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \dots & y_{KN_{rx}} \end{bmatrix}$ denote the channel output symbol vector at the receiver, in which $J, K, M,$ and N_{rx} represent the number of users, the NOMA signature length, the order of modulation, and the number of receive antennas, respectively. Assuming equi-probable input constellation points and the high-dimensional

[19] The average mutual information is defined as the weighted sum of the mutual information between each pair of the input and output events x_i and y_j. The average mutual information is a measure of independence between the two random variables X and Y. In mathematical terms $I(X;Y) = E[I(x_i;y_i)] = \sum_i \sum_j p(x_i,y_j) I(x_i;y_j)$ and $I(x_i;y_i) = \log_2 P(x_i,y_i)/P(x_i)P(y_i)$.

constellation set given by $\Omega = \{\omega_1, \omega_2, \ldots, \omega_{2^{MJ}-1}\}$, the average mutual information in the multi-user case can be written as [51]:

$$I(\mathbf{x};\mathbf{y}) = \frac{I(\mathbf{x};\mathbf{y}|\mathbf{H})}{K} = \frac{J\log_2 M}{K} - \frac{1}{K}E_{\mathbf{x},\mathbf{y},\mathbf{H}}\left\{\log_2\left[\frac{\sum_{\omega\in\Omega}P(\mathbf{y}|\mathbf{x}=\omega,\mathbf{H})}{P(\mathbf{y}|\mathbf{x},\mathbf{H})}\right]\right\}$$

A Monte Carlo simulation can be used to calculate the expectation function in the above equation. Assuming a six-user uplink multiple access channel, a tapped delay line TDL-A-30 ns channel $\mathbf{H}$ and ideal channel estimation, multi-user average mutual information of the NOMA schemes under consideration in 3GPP for Rel-16 and OFDMA have been calculated and compared in Fig. 3.37. In this analysis, the overloading factor is 150%, spectral efficiency per user is 0.25 bps/Hz, and the sum-rate is 1.5 bps/Hz. The SNR is defined as the ratio of average total received multi-users' power to the noise power at each receive antenna for given bandwidth. Multiple UEs are assumed to share the same six physical resource blocks. The number of users, SE per user, and transmission bandwidth are identical for both NOMA and OFDMA schemes to ensure fairness.

Fig. 3.37 further compares the BLER performance of the NOMA and OFDMA schemes for six and eight UEs, respectively. LTE turbo code is used for the channel coding and QPSK 1/2 for NOMA QPSK 3/4 is used for OFDMA, the overloading factor is 150%, spectral efficiency per user is 0.25 bps/Hz, and the sum-rate is 1.5 bps/Hz. In theory, the multi-user average mutual information analysis suggests that NOMA schemes provide higher capacity relative to OFDMA for given achievable sum-rate. In addition the, coding-based NOMA schemes (SCMA) have some performance advantage over other schemes. When the number of UEs increases the OFDMA system needs to use higher order modulation which would

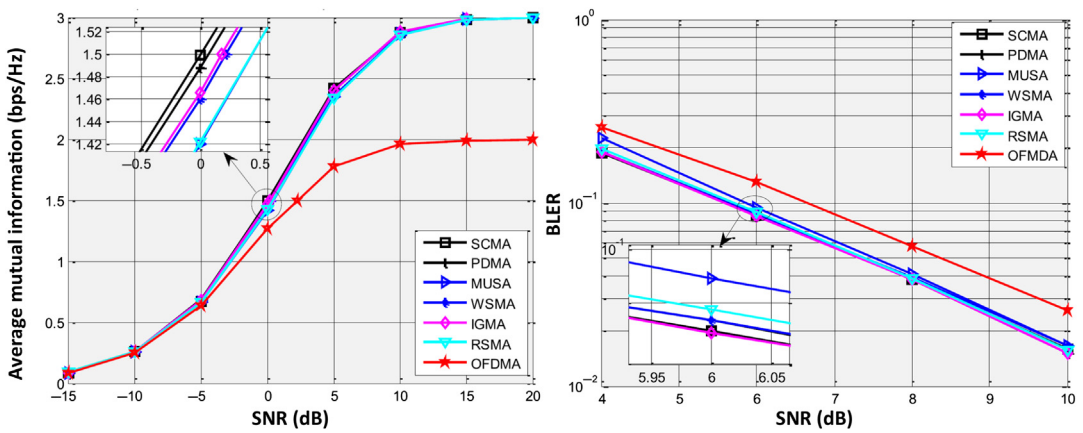


Figure 3.37
Comparison of average mutual information and BLER of NOMA and OFDMA schemes [51].

suffer from performance loss, while NOMA schemes with low order modulation can take advantage of superposition coding for higher overloading factor with slight performance degradation. The BLER performance advantage of NOMA schemes over OFDMA grows with the increase of the number of UEs [51].

3.4 Duplex Schemes

One of the key elements of any radio communication system is the way in which radio communications are maintained in the downlink and uplink. For cellular systems, it is necessary to enable simultaneous transmission of data in both directions, which creates a number of constraints on the schemes that may be used to control over the air transmission. As a result, the choice of duplex scheme becomes the basic part of the overall specification for the cellular or any radio communications system. The term duplex refers to the bidirectional communications between two devices. When unpaired spectrum or alternatively the same RF carrier is used for downlink and uplink communications, the transmit/receive functions are time-multiplexed. When paired spectrum or alternatively two RF carriers are used for downlink and uplink communications, the transmit/receive functions are frequency-multiplexed.

3.4.1 Frequency and Time Division Duplex Schemes

The Frequency Division Duplex is a duplex scheme in which uplink and downlink transmissions occur simultaneously using different frequencies. The downlink and uplink frequencies are separated by sufficiently large frequency offset. For the FDD scheme to properly operate, it is necessary that the frequency separation, that is, channel separation between the transmission and reception frequencies, to be sufficient in order to prevent the receiver blocking due to high-power transmitter signal. The receiver blocking is an important issue in FDD schemes and often highly selective filters may be required. For cellular systems using FDD, filters are required in the base station and the user terminal to ensure sufficient isolation of the transmitter signal without desensitizing the receiver. While implementation cost is not a significant constraint for the base stations, placing a filter in the user terminal involves higher complexity and cost. The use of an FDD system does enable simultaneous transmission and reception of signals. However, two RF channels are required, which in some cases may not be the efficient use of the available spectrum. The spectrum used for FDD systems is allocated by the regulatory bodies. Since there is a frequency separation between the uplink and downlink directions, it is not typically possible to reallocate spectrum to change the balance between the capacity of the uplink and downlink directions, if the capacity requirements for each direction vary over time.

The Time Division Duplex is a duplex scheme where uplink and downlink transmissions occur at different times but may share the same frequency. In other words, the downlink and uplink transmissions are multiplexed in time and are not concurrent. While FDD transmissions require a large frequency separation between the transmitter and receiver frequencies, TDD schemes require a guard time or guard interval between transmission and reception. This gap must be sufficient to allow the signals traveling from the remote transmitter to arrive before a transmission is started and the receiver is shut down. Although this delay is relatively short, switching between transmission and reception several times in a second, even a small guard time can reduce the spectral efficiency of the system since a percentage of time must be used for the guard interval. For small-sized cells, for example, up to one mile, the guard interval is typically small and acceptable. However, for large cell sizes, it may become an issue and may introduce significant overhead.

FDD has been the dominating duplex scheme since the beginning of the mobile communication era. In the 5G era, FDD will remain the main duplex scheme for lower frequency bands; however, for higher frequency bands, especially above 10 GHz and targeting very dense deployments, TDD will play more important role. In very dense deployments with low-power nodes, the TDD-specific interference scenarios (direct base station-to-base-station and device-to-device interference) will be similar to the base-station-to-device and device-to-base-station interference that also occurs in FDD schemes. Furthermore, for the dynamic traffic variations expected in very dense deployments, the ability to dynamically assign transmission resources (e.g., time slots) to different transmission directions may allow more efficient utilization of the available spectrum. To reach its full potential, 5G will allow for very flexible and dynamic assignment of TDD transmission resources. This contrasts with current TDD-based mobile technologies, including TD-LTE, for which there are restrictions on the downlink/uplink configurations; thus there typically exist assumptions about using the same configuration for neighboring cells and between neighboring operators.

The guard interval required for TDD will comprise two main elements: (1) A time-allowance for the propagation delay for any transmission from a remote transmitter to arrive at the receiver. This will depend upon the distances involved (i.e., cell radius) and (2) a time-allowance for the transceiver to switch from receive-to-transmit mode. The switching times can vary considerably depending on the implementation but can take a few microseconds. As a result, TDD is not normally suitable for use over very large cell sizes as the guard time increases and the spectral efficiency decreases.

It is often found that traffic in both directions is not balanced. Typically, there is more data transmitted in the downlink direction of a cellular system. This means that the capacity should be ideally greater in the downlink direction. Using a TDD system, it is possible to

change the capacity in either direction simply by changing the number of time slots allocated to each direction. This is often dynamically configurable so it can be adapted to meet the demand. A further aspect to be noted with TDD transmission is the latency. Since data may not be able to be routed immediately to the transmission chain as a result of the time multiplexing between transmit and receive circuitry, there will be a small delay between the data being generated and being actually transmitted. Typically, this may be a few milliseconds depending on the frame timing. Both TDD and FDD have their advantages and each can be used in different deployment scenarios. Before deciding on a particular type of duplex scheme, it is necessary to analyze the advantages and disadvantages of each duplex mode. Table 3.7 summarizes and compares the relative attributes of TDD and FDD systems.

The new radio can operate in paired and unpaired spectrum using a common frame structure unlike LTE where two different frame structures were used, which was later expanded to three for support of unlicensed operation introduced in 3GPP Rel-13. The basic NR frame structure is designed such that it can support both half-duplex and full-duplex operation. In half duplex the device cannot transmit and receive at the same time.

A major issue in limiting the capacity of non-cooperative cellular massive MIMO networks operating in TDD mode is the pilot contamination. The rise of asymmetric uplink/downlink traffic patterns necessitates that the ratio between downlink/uplink traffic changes over the time; thus the static paired spectrum for downlink/uplink is not efficient for supporting such dynamic asymmetric traffic, particularly in UDNs. Flexible duplex can better adapt to dynamic asymmetric traffic. With flexible duplex, the uplink spectrum defined in FDD systems can be reallocated for downlink transmission with high flexibility. Considering the potential cross-link interference, that is, downlink to uplink and uplink to downlink, the transmission power for downlink transmission on the uplink spectrum should be constrained to a relatively low. Flexible duplex can be applicable to small cells with low transmission power and relay base stations.

3.4.2 Half-Duplex and Flexible-Duplex Schemes

The LTE and NR support TDD and FDD schemes with a great extent of commonality in the baseband processing. In order to reduce the implementation complexity and cost of FDD terminals and further to increase the reuse of baseband functional elements, a half-duplex FDD (H-FDD) operation is supported where the downlink and uplink transmissions are not simultaneous, but occur in two different frequencies. A classic H-FDD operation does not efficiently utilize the radio resources on the downlink and uplink RF carriers. The complementary grouping and scheduling of users would allow efficient use of downlink and uplink resources in an H-FDD operation. For H-FDD operation, a guard period is virtually

Table 3.7: Comparison of time division duplex (TDD) and frequency division duplex (FDD) attributes.

Attribute	TDD	FDD
Paired spectrum	Does not require paired spectrum as both transmit and receive occur on the same channel	Requires paired spectrum with sufficient frequency separation to allow simultaneous transmission and reception
Hardware cost	Lower cost as no diplexer[a] is needed to isolate the transmitter and receiver	Diplexer is needed and the implementation cost is higher
Channel reciprocity	Channel propagation is the same in both directions which enables estimation of the downlink channel from the uplink channel	Channel characteristics are different in both directions as a result of the use of different frequencies
UL/DL asymmetry	It is possible to dynamically change the UL and DL ratio based on the traffic volume in each direction	UL/DL bandwidths are determined by frequency allocation designated by the regulatory authorities. It is therefore not possible to dynamically adapt to the traffic volume
Guard period/ frequency separation	Guard period required to ensure uplink and downlink transmissions do not interfere. Large guard period will limit the capacity. Larger guard period normally required if distances are increased to accommodate larger propagation delays. Note that a guard band in frequency domain is required to suppress interference to adjacent bands.	Frequency separation is required to provide sufficient isolation between uplink and downlink. However, large frequency separation does not impact the capacity. Note that a guard band in frequency domain is required to suppress interference to adjacent bands.
Discontinuous transmission	Discontinuous transmission is required to allow both uplink and downlink transmissions. This can degrade the performance of the RF power amplifier in the transmitter	Continuous transmission is possible
Switching point synchronization	Base stations are required to be synchronized with respect to the uplink and downlink transmission times. If neighboring base stations use different uplink and downlink assignments and share the same channel, then interference may occur between cells.	Not applicable

[a]A diplexer is a passive device that implements frequency domain multiplexing. Two ports are multiplexed onto a third port. The signals on each input ports occupy nonoverlapping frequency bands. Consequently, the signals on the input ports can coexist on the output port without interfering with each other. On the other hand, a duplexer is a device that allows bidirectional communication over a single channel. In radar and radio communications systems the duplexer isolates the receiver from the transmitter while permitting them to share a common antenna. Most radio repeater systems include a duplexer. Duplexers are designed for operation in the frequency band used by the receiver and transmitter and must be capable of handling the output power of the transmitter. They must provide sufficient isolation between transmitter and receiver to prevent receiver desensitization [30].

created by the UE by not receiving the last OFDM symbol(s) of a downlink subframe immediately preceding an uplink subframe in which the UE is active. The length of the guard period is the sum of the maximum round-trip propagation delay in the cell, transmit-to-receive and receive-to-transmit switching delay at the UE.

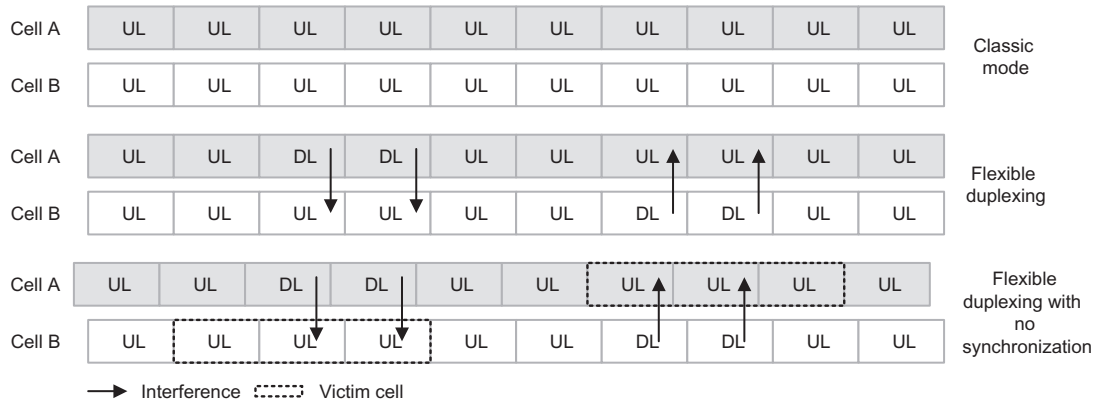


Figure 3.38
Examples of flexible duplexing schemes in TDD mode.

The choice of duplex scheme is typically determined by the spectrum allocation. For lower frequency bands, allocations are often paired, implying FDD mode is dominant. However, in higher frequency bands, unpaired spectrum allocations are increasingly common; thus TDD mode is used. 3GPP NR supports both duplexing methods. However, unlike LTE where the TDD uplink-downlink allocation does not change over time, NR supports dynamic TDD as a key technology component. In dynamic TDD, parts of slot can be dynamically allocated to either uplink or downlink based on the scheduler decision. This enables network adaptation to traffic variation, which is particularly common in dense deployments with a relatively small number of users per base station. The device follows any scheduling decisions, and it is up to the network scheduler, if necessary, to coordinate the scheduling decisions between neighboring sites to avoid inter-cell interference. There is also a possibility to semi-statically configure the transmission direction of some of the slots, a feature that can allow reduced device energy consumption as it is not necessary to monitor downlink control channels in slots that are a priori known to be reserved for uplink usage (Fig. 3.38).

3.4.3 Full-Duplex Schemes

The term full-duplex was traditionally used to describe a simultaneous bidirectional communication, in contrast to half-duplex, which assumed time division duplexing. However, in recent years, the term has carried a new concept and that is when a device can transmit and receive at the same time and over the same carrier frequency. Some authors refer to the latter scheme as in-band full duplex in the literature to distinguish this new concept from its traditional usage. It is intuitive that enabling wireless devices to operate in full-duplex mode offers the potential to double the spectral efficiency, considering that traditional approaches for increasing spectral efficiency such as adaptive coding and modulation, MIMO and smart antennas have almost reached their maximum limits. In addition, full-duplex scheme

improves the reliability and flexibility of dynamic spectrum allocation in cognitive radio networks and enables the small cells to reuse radio resources simultaneously for access and backhaul [29].

The main challenge for a full-duplex radio is self-interference and how to manage and suppress it. Self-interference or transmitter leakage was studied earlier and refers to the signal that leaks from the device transmitter to its own receiver. In general, the transmitter signal is about 100 dB stronger than the reference sensitivity level of the receiver. A considerable part of this transmitted signal leaks into the receiver chain, causing serious issues in decoding the desired signal, which could be considered noisy, with a dramatically affected SNR. To achieve the best performance of a full-duplex system, the self-interference signal must be suppressed to reach the receiver's noise floor. The self-interference may be originated from the linear components and the RF carrier itself, which is attenuated and reflected from the environment. The received distortion can be modeled as a linear combination of different delayed copies of the original carrier. It can be further generated by nonlinear components because the imperfect radio circuits typically create third and fifth intermodulation products of the transmit signal. These higher order intermodulation terms have significant frequency content at frequencies close to the transmitted frequencies, which directly correspond to other harmonics. The self-interference can also be caused by the transmitter noise, which appears as an increase of about 50 dB over receiver noise floor. Due to random nature of this component, it can only be canceled by subtracting the appropriately weighted transmitter signal sampled in the analog domain from the received signal. For narrowband systems, the self-interference channel can be modeled as gain and delay functions, whereas wideband systems require a more complex model, because the reflected-path self-interference channel is often frequency-selective as a result of multipath propagation.

In general, one can model the process in the digital domain that is valid for both the narrowband and wideband scenarios as $r(n) = r_d(n) + w_r(n) + r_{DSI}(n) + r_{SSI}(n)$ where $r(n)$ is the total received complex-valued baseband samples, $r_d(n)$ is the desired signal from the remote node, $r_{DSI}(n)$ denotes the complex-valued samples caused by the direct self-interference component signal between the transmit and receive antennas in case of two antennas, or leaked signal in the circulator in case of one antenna, $r_{SSI}(n)$ represents the complex-valued samples caused by scattered self-interference components, and $w_r(n)$ denotes additive white Gaussian noise. A circulator is a passive three-port device in which an RF signal entering any port is only transmitted to the next port in rotation. Both direct and scattered self-interference can be represented as a combination of linear and nonlinear components. The suppression in the propagation domain can mitigate both linear and nonlinear self-interference at the same time and with the same isolation value. Meanwhile, cancellation techniques in the analog and digital domains have different performances for the two components. The self-interference cancellation is implemented in three domains: propagation, analog, and digital. Since none of these domains can meet the required cancellation

requirements, hybrid solutions have been proposed in the literature [40,41]. The primary role of self-interference cancellation in the propagation and analog domains is to avoid the saturation of the receiver due to the high power of the self-interference signal as this power exceeds the ADC dynamic range and limits its precision after conversion because the desired signal is much weaker than the self-interference.

Self-interference comprises several components with different characteristics depending on the specifics of the full-duplex system implementation, such as the number of antennas, the characteristics of the RF components and the environment. The constituents of the self-interference can be classified by linearity. Linear components involve multipath propagation between the transmit/receive antennas. For a single-antenna system, linear components include the leakage of the circulator or the reflections from impedance mismatch. Components of RF circuits, such as attenuators and delay lines, are also modeled as linear systems for analog self-interference cancelation. The linear components of self-interference can be removed by the existing channel estimation methods, as in most conventional wireless communication systems. Nonlinear components are usually created by PAs in transmitters and low-noise amplifiers in receivers. The nonlinearity of the PA is generated because the power of the output signal is saturated for the high-power input signal, which worsens for modulation schemes with high PAPR such as OFDM and wideband CDMA. Intermodulation distortion, caused by the nonlinearity, interferes with the linear model of self-interference. The intermodulation can be theoretically calculated by a Volterra series[20] or approximated by a Taylor series. Since the even-ordered terms are out-of-band, the Taylor series would include only odd-ordered terms. Other RF imperfections/impairments, for example, I/Q imbalance, phase noise and transmitter noise can occur in the transmitter side. The I/Q imbalance occurs when there is mismatch between the gain and phase of the two sinusoidal signals, which deteriorates the baseband transmitter signal. The imperfection of the local oscillator can also degrade the linearity of the transmitter signal. In general, most of the impacts of the oscillator impairment is noticeable in random deviations in the output frequency, which can be modeled as phase noise. Transmitter noise also includes

[20] Volterra series is a mathematical model for nonlinear behavior of systems in which the output of the nonlinear system always depends on the input. This provides the ability to capture the memory effect of devices. In mathematics, Volterra series denotes functional expansion of a dynamic, nonlinear, time-invariant function. The Volterra series, which is used to prove the Volterra theorem, is an infinite sum of multidimensional convolutional integrals. A continuous time-invariant system with $x(t)$ as input and $y(t)$ as output can be expanded in Volterra series as $y(t) = h_0 + \sum_{n=1}^{N} \int_a^b \dots \int_a^b h_n(\tau_1, \dots, \tau_n) \prod_{m=1}^{n} x(t - \tau_m) d\tau_m$ where the constant term h_0 on the right-hand side is often set to zero by suitable choice of output level y. The function $h_n(\tau_1, \dots, \tau_n)$ is called the nth order Volterra kernel. It can be regarded as a higher-order impulse response of the system. If N is finite, the series can be truncated. If $a, b,$ and$N$ are finite, the series is called doubly finite. Since in any physically realizable system, the output can only depend on previous values of the input, the kernels $h_n(t_1, \dots, t_n)$ will be zero, if any of the variables $t_1, t_2, \dots, t_n$ are negative. The integrals may then be written over the half range from zero to infinity. Therefore, if the operator is causal $a \geq 0$.

thermal noise which is typically generated by RF components. Using the estimated linear wireless channel, this method can mitigate transmitter impairments. Unlike other SIC methods, the distortion of the transmitter signal by PA nonlinearity or by I/Q imbalance is obtained directly by the auxiliary receive chain.

The specifications of a full-duplex system can be classified into three categories: main specifications, ADC specifications, and self-interference specifications. While the residual self-interference power P_{RSI} is higher than the receiver noise floor level, the signal to self-interference plus noise ratio in a full-duplex system is lower than the SNR of a half-duplex system receiver. This means that the maximum efficiency of full-duplex cannot be achieved. In general, self-interference cancellation solutions are a combination of several techniques to help meet the system requirements. Fig. 3.39 provides an example, showing the average performance value achieved in each domain.

In case of a shared transmit/receive antenna system, the suppression is performed using a three-port RF circulator. The ferrite within the device can be considered as a propagation domain. Achievable isolation by the circulator is between 15 and 30 dB, and in the case of wideband operation, the maximum value would decrease. In a separate-antenna system, several self-interference cancellation techniques can be used. The two transmit antennas can be placed at distances d and $d + \lambda/2$ away from the receive antenna. Separating the two transmitters by half a wavelength causes their signals to cancel one another. For narrowband signals, this technique is experimentally proved to be sufficiently robust; however, the suppression drastically decreases in case of wideband signals. The antenna directionality isolates the receiving antenna from the interfering signals of the transmitting antenna;

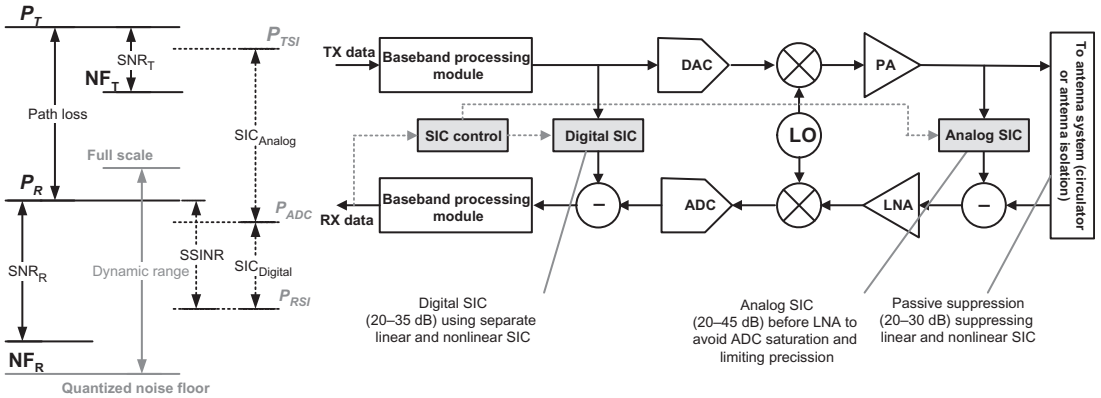


Figure 3.39

Example Wi-Fi or LTE signals, which are typically transmitted at an average power of 23 dBm (200 mW). The thermal noise level at 20 MHz bandwidth is about −101 dBm (−174 dBm/Hz + 73 dB) [40,41].

however, such an approach would not work for point-to-point full-duplex scenarios. Electromagnetic shielding can enhance the isolation between transmit and receive antennas. Nevertheless, one disadvantage is that the shielding affects the far-field antenna patterns because it prevents the antenna from transmitting to/receiving from the shielding direction; thus it is only relevant to the case of directional antennas. Self-interference can be mitigated using orthogonal polarization between the transmit/receive antennas, achieving about 10−20 dB isolation in an anechoic chamber and 6−9 dB in a reflective room at 2.4 GHz. The dual-polarized antenna can suppress self-interference by transmitting and receiving through orthogonal polarization. The isolation characteristic of a dual-polarized antenna can be expressed by its XPD factor. In transmit mode, XPD is the proportion of the signal that is transmitted in the orthogonal polarization to the desired direction. In receive mode, it is the antenna's ability to maintain the incident signal's polarization purity. For example, if a perfectly vertically polarized signal (containing no horizontal component) was incident upon a single polarized receive antenna, the electrical and mechanical imperfections would introduce a small amount of ellipticity to the polarization of the signal. The signal can be thought of as having both vertical and horizontal components. The ratio of the resulting horizontal to vertical components is the defined as XPD.

Active self-interference cancellation techniques in the analog domain generate a replica of the transmit signal and then adjust it to match the self-interference channel, making the replica similar to the self-interference signal in order to subtract it from the total received signal. This replica can be generated either in the analog domain or in the digital domain before the DAC. The self-interference cancellation signal is performed in the same domain from which it was created; thus no additional ADC/DAC is required. Replication of the transmission signal in the analog domain can be achieved by tapping the transmitter chain, using a power splitter, or using a balanced-to-unbalanced circuit in the case of two separate antennas. Experiments show the practical benefits of the latter approach relative to phase shifter, notably the flatter response within a wide frequency band. After creating an exact negative replication of the RF reference signal, the replica is adjusted by delay and attenuation elements to match the self-interference [40,41] (Fig. 3.40).

It can be observed in the experiments that the nonlinear components of self-interference can be 80 dB higher than the receiver noise floor. A large fraction of this component may be eliminated along with the linear self-interference by self-interference cancellation techniques in the analog and propagation domains, but the residual nonlinear self-interference in the digital domain, which amounts to 10−20 dB, needs to be canceled. In general, nonlinear self-interference cancellation methods are added to the linear methods to achieve optimal performance. A generic model for approximation of the nonlinear function is based on a Taylor series, where the output transmitted signal is represented as $y(n) = \sum_m a_m x_m^p(n)$ where $x_m^p(n)$ is the ideal passband analog signal for the digital representation of $x(t)$. It can

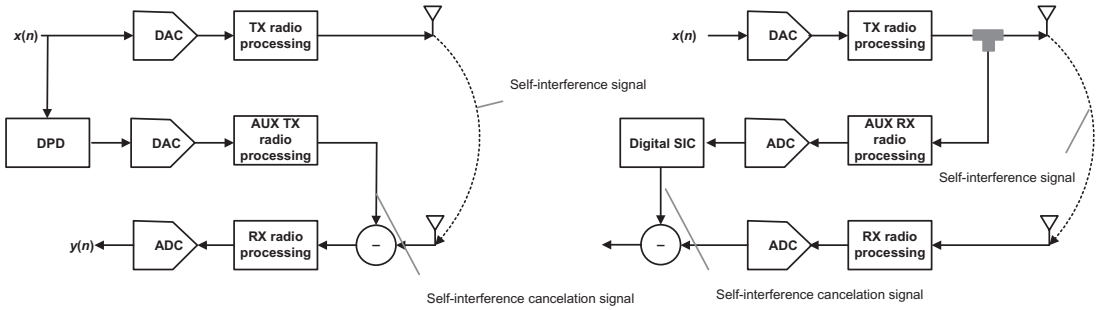


Figure 3.40
Auxiliary transmit/receive chain in full-duplex systems [40,41].

be shown that for practical wireless systems, only the odd orders of the polynomial contribute to the in-band distortion. Furthermore, only a limited number of odd orders contribute to the distortion, and higher orders can be neglected. The nonlinearity is typically characterized by the third-order intercept point, which is defined as the point at which the power of the third harmonic is equal to the power of the first harmonic. Another source of impairment is IQ imbalance which is caused by the gain and phase mismatches between I and Q branches of the transmitter and receiver chains. This imbalance results in the complex conjugate of the ideal signal to be superimposed with some level of attenuation. Thus the output of an imperfect IQ mixer is a transformation of an input signal $x(t)$ where both direct and conjugated signals are filtered and then summed together. The IQ imbalance can be modeled and compensated using widely linear filters.[21]

The results of studies suggest that the oscillator phase noise is one of the main self-interference cancellation challenges that limit the performance of full-duplex systems. It was assumed that when transmitter and receiver use a common local oscillator, the level of phase noise would remain at a tolerable level. However, this consideration is not always valid, especially in the case of OFDM systems. The studies show that with a phase noise variance between 0.4 and 1.0 degrees, the reduction in self-interference cancellation performance is about 20−25 dB for OFDM systems. This can be explained by the phase noise causing two effects: common phase error and inter-carrier interference. The former may have acceptable levels as previously assumed, but the latter stimulates an enhancement in

[21] Widely linear filters augment the data vector with its conjugate, thus providing the complete second-order statistical information when computed using the minimum mean square error cost function. Widely linear filters have been proposed for applications such as interference cancelation, demodulation, and equalization for direct sequence code-division-multiple-access systems, and array receivers. A widely linear filter forms the estimate of a desired sequence $d(n)$ through the inner product $y_{WL}(n) = \mathbf{w}^H \overline{\mathbf{x}}(n)$ where the weight vector $\mathbf{w} = [w_0 \;\; w_1 \;\; \ldots \;\; w_{2N-1}]^T$ has double dimension compared to the linear filter and $\overline{\mathbf{x}}(n) = [\mathbf{x}(n)\mathbf{x}^*(n)]^T$ with the definition $\mathbf{x}(n) = \begin{bmatrix} x(n) & x(n-1) & \cdots & x(n-N+1) \end{bmatrix}^T$.

self-interference cancellation performance, which is achieved by consecutively estimating and suppressing the ICI.

The goal of digital self-interference cancellation is to remove the residual self-interference after analog self-interference cancellation especially that originated from NLoS reflections. Various signal processing and calculations performed after ADC for the purpose of self-interference cancellation are classified as digital self-interference cancellation. Since they operate in the digital domain, the baseband equivalent models of self-interference should be determined before calculation. The models include linear and nonlinear self-interference models. Once the self-interference signal is modeled as a function of the transmitted signal, it can be estimated from the transmitted and received signals. The residual self-interference is then reconstructed as the output of the function, and subtracted from the received signal [40,41].

3.5 Operating Frequency Bands

The new radio has been developed to operate in two distinct frequency regions: sub-6 GHz referred to as FR1, and mmWave referred to as FR2 in 3GPP radio specifications [9]. In each band, the new radio may operate in either FDD or TDD duplex modes. A band may be a supplementary downlink (SDL) or supplementary uplink (SUL) band used to provide additional capacity in the respective direction. Given the wide range of frequencies which 5G NR is required to support, three subcarrier-spacings (15, 30, and 60 kHz) are identified for operation in FR1 for sub-6 GHz bands; and two subcarrier-spacings (60 and 120 kHz) are designated for operation in FR2 for mmWave bands. Channel bandwidths vary based on SCS, with many supporting 5 MHz channel width with an SCS of 15 kHz. The FR1 covers a frequency range from 450 MHz to 6 GHz (this limit may be extended to 7.125 GHz in Rel-16) whereas FR2 spans a frequency range from 24.25 to 52.6 GHz. Note that frequency bands beyond 52.6 GHz will be supported in future releases of 3GPP once ratified by WRC-19 in late 2019. The NR operating bands have been shown in Table 3.8.

The gNB channel bandwidth supports a single NR RF carrier in the uplink or downlink. Different UE channel bandwidths may be supported within the same spectrum for transmitting to and receiving from UEs connected to the gNB. The assignment of the UE channel bandwidth is flexible and within the gNB channel bandwidth. The NR base station is able to transmit to and/or receive from one or more UE bandwidth parts (BWPs) that are smaller than or equal to the number of carrier resource blocks on the RF carrier [9]. The number of physical resource blocks configured in any channel bandwidth ensures that the required minimum guard band between neighboring channels is satisfied. The relationship between channel bandwidth, guard band and transmission bandwidth is shown in Fig. 3.41 for a single NR channel. The maximum number of resource blocks for each channel bandwidth and each frequency range is given in Table 3.9. Since the physical resource blocks always

Table 3.8: NR operating bands in frequency range (FR) 1 and FR2 [9].

NR Frequency Range	NR Operating Band	Uplink Operating Bands		Downlink Operating Bands		Duplex Mode
		F_{UL_Low}–F_{UL_High} (MHz)	Total Bandwidth (MHz)	F_{DL_Low}–F_{DL_High} (MHz)	Total Bandwidth (MHz)	
FR1	*n*1	1920–1980	60	2110–2170	60	FDD
	*n*2	1850–1910	60	1930–1990	60	FDD
	*n*3	1710 -1785	75	1805–1880	75	FDD
	*n*5	824–849	25	869–894	25	FDD
	*n*7	2500–2570	70	2620–2690	70	FDD
	*n*8	880–915	35	925–960	35	FDD
	*n*20	832–862	30	791–821	30	FDD
	*n*28	703–748	45	758–803	45	FDD
	*n*38	2570–2620	50	2570–2620	50	TDD
	*n*41	2496–2690	194	2496–2690	194	TDD
	*n*50	1432–1517	85	1432–1517	85	TDD
	*n*51	1427–1432	5	1427–1432	5	TDD
	*n*66	1710–1780	70	2110–2200	90	FDD
	*n*70	1695–1710	15	1995–2020	25	FDD
	*n*71	663–698	35	617–652	35	FDD
	*n*74	1427–1470	43	1475–1518	43	FDD
	*n*75	N/A		1432–1517	85	SDL
	*n*76	N/A		1427–1432	5	SDL
	*n*78	3300–3800	500	3300–3800	500	TDD
	*n*77	3300–4200	900	3300–4200	900	TDD
	*n*79	4400–5000	600	4400–5000	600	TDD
	*n*80	1710–1785	75	N/A		SUL
	*n*81	880–915	35	N/A		SUL
	*n*82	832–862	30	N/A		SUL
	*n*83	703–748	45	N/A		SUL
	*n*84	1920–1980	60	N/A		SUL
FR2	*n*257	26,500–29,500	3000	26,500–29,500	3000	TDD
	*n*258	24,250–27,500	3260	24,250–27,500	3260	TDD
	*n*260	37,000–40,000	3000	37,000–40,000	3000	TDD

consists of 12 subcarriers, the maximum number of resource blocks for each channel bandwidth depends on the subcarrier spacing.

The spacing between carriers will depend on the deployment scenario, the size of the frequency block available and the channel bandwidths. The nominal channel spacing df_s between two adjacent NR carriers with 100 kHz channel raster is defined as $df_s = (\mathrm{BW}_{ch1} + \mathrm{BW}_{ch2})/2$ where BW_{ch1} and BW_{ch2} denote the channel bandwidths of the two respective NR carriers. The channel spacing can be adjusted depending on the channel raster to optimize performance in a particular deployment scenario. For NR carriers in FR1

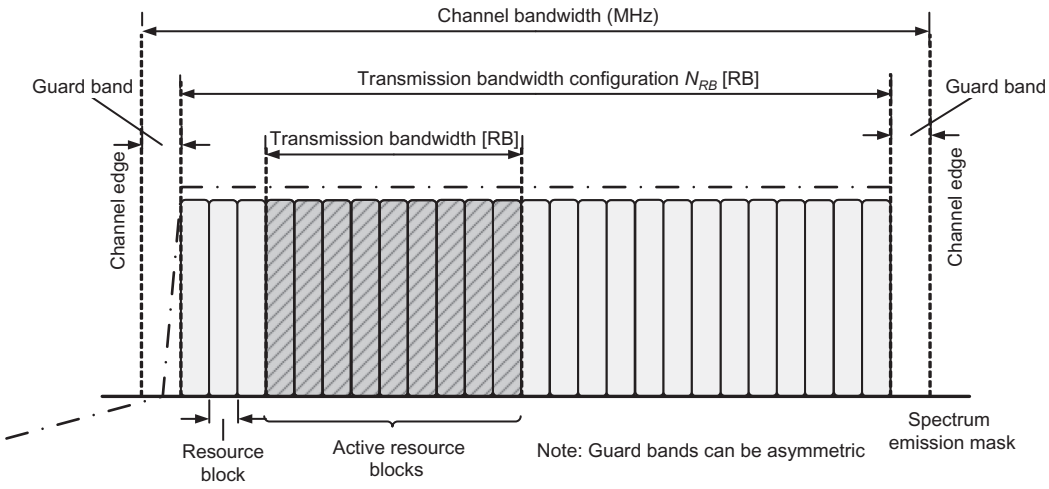


Figure 3.41
Definition of channel and transmission bandwidth configuration for one NR channel [9].

Table 3.9: Maximum transmission bandwidth configuration in N_{RB} for frequency ranges 1 and 2 [9].

		FR1													FR2			
		Bandwidth (MHz)													Bandwidth (MHz)			
Subcarrier Spacing (SCS) (kHz)		**5**	**10**	**15**	**20**	**25**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**	**50**	**100**	**200**	**400**
15	Number	25	52	79	106	133	160	216	270	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
30	of	11	24	38	51	65	78	106	133	162	189	217	245	273	N/A	N/A	N/A	N/A
60	resource	N/A	11	18	24	31	38	51	65	79	93	107	121	135	66	132	264	N/A
120	blocks N_{RB}	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	32	66	132	264

operating bands with 15 kHz channel raster, the nominal channel spacing is defined as $df_s = (\mathrm{BW}_{ch1} + \mathrm{BW}_{ch2})/2 \pm [5\mathrm{kHz}, 0]$. Furthermore, for NR carriers in FR2 operating bands with 60 kHz channel raster, the nominal channel spacing is defined as $df_s = (\mathrm{BW}_{ch1} + \mathrm{BW}_{ch2})/2 \pm [20\mathrm{kHz}, 0]$ [9].

In the case that multiple numerologies are multiplexed over the same symbol, the minimum guard band on each side of the carrier is the guard band applied at the configured gNB channel bandwidth for the numerology that is transmitted/received immediately adjacent to the guard band. Nevertheless, if multiple numerologies are multiplexed on the same symbol and the gNB channel bandwidth is wider than 50 MHz when operating in FR1, the guard band applied adjacent to 15 kHz SCS is the same as the guard band

defined for 30 kHz SCS for the same channel bandwidth. If multiple numerologies are multiplexed on the same symbol and the gNB channel bandwidth is wider than 200 MHz when operating in FR2, the guard band applied adjacent to 60 kHz SCS is the same as the guard band defined for 120 kHz SCS for the same channel bandwidth. For each numerology, the starting point of its transmission bandwidth configuration on the common resource block (CRB) grid for a given channel bandwidth is indicated by an offset to reference point A in the unit of the numerology while the indicated transmission bandwidth configuration must fulfill the minimum guard band requirements [9].

The channel raster defines a subset of RF reference frequencies that can be used to identify the uplink and downlink RF channel positions, where the RF reference frequency corresponding to an RF channel is mapped to a resource element on that carrier. A global frequency raster is further defined for all frequencies from 0 to 100 GHz and is used to define the set of RF reference frequencies F_{REF} that are used for signaling the location of RF channels and synchronization signal blocks. The granularity of the global frequency raster is defined as ΔF_{global}. For each operating band, a subset of frequencies from the global frequency raster are applicable for that band and form a channel raster for that band with a granularity $\Delta F_{raster} \geq \Delta F_{global}$. The channel raster is mapped to physical resource block $n_{PRB} = \lfloor N_{RB}/2 \rfloor$ with resource element index $k = 0$ or 6 depending on whether $N_{RB} \bmod 2 = 0$ or 1, respectively [9].

The RF reference frequency in the uplink and downlink is identified by the NR absolute radio frequency channel number (NR-ARFCN) in the range [0...3279167] on the global frequency raster. The relationship between the NR-ARFCN and the RF reference frequency F_{REF} in MHz for the downlink and uplink is given as $F_{REF} = F_{REF-offset} + \Delta F_{raster}\left(N_{REF} - N_{REF-offset}\right)$, where $F_{REF-offset}$ and $N_{REF-offset}$ are given in Table 3.10 and N_{REF} represents the NR-ARFCN [9].

For the supplementary bands and bands $(n1, n2, n3, n5, n7, n8, n20, n28, n66, n71)$ (see Table 3.8), the reference frequency is defined as $F_{REF\text{-}shift} = F_{REF} + \Delta_{shift}$ where $\Delta_{shift} = 0$ or 7.5kHz is signaled by the network.

A channel raster of 100 kHz is used for some NR operating bands, in which $\Delta F_{raster} = 20\Delta F_{global}$; thus every 20th NR-ARFCN within the operating band can be used for the channel raster with the step size of 20. For NR operating bands below 3 GHz with 15 kHz channel raster $\Delta F_{raster} = n_{step}\Delta F_{global}$. In this case, every $n_{step} \in \{3, 6\}$ NR-ARFCN within the operating band is a candidate channel raster. There are NR operating bands above 3 GHz where the channel raster is either 15 or 60 kHz. In that case, $\Delta F_{raster} = n_{step}\Delta F_{global}$ and every $n_{step} \in \{1, 2\}$ NR-ARFCN within the operating band is a candidate channel raster.

Table 3.10: NR absolute radio frequency channel number parameters for the global frequency raster [9].

Frequency Range (MHz)	ΔF_{global} (kHz)	$F_{REF-offset}$ (MHz)	$F_{REF-offset}$	N_{REF} Range
0–3000	5	0	0	0–599,999
3000–24,250	15	3000	600,000	600,000–2,016,666
24,250–100,000	60	24,250	2,016,667	2,016,667–3,279,167

Table 3.11: Global synchronization channel number (GSCN) parameters for the global frequency raster [9].

Frequency (MHz)	Synchronization Block Frequency Position SS_{REF}	GSCN	Range of GSCN
0–3000	1200 N kHz + 50 M kHz; $N = 1{:}2499$, $M \in \{1,3,5\}$	$3N + (M-3)/2$	2–7498
3000–24,250	3000 MHz + 1.44 N MHz; $N = 0{:}14{,}756$	$7499 + N$	7499–22,255
24,250–100,000	24250.08 MHz + 17.28N MHz; $N = 0{:}4383$	$22,256 + N$	22,256–26,639

The synchronization raster signifies the frequency positions of the synchronization block that can be used by a UE for frequency acquisition when the UE has not received an explicit signaling indicating the synchronization block position. A global synchronization raster is defined for all frequencies. The frequency position of the synchronization block is defined by parameter SS_{REF} with a corresponding global synchronization channel number (GSCN). The GSCNs are numbered in increasing frequency order as shown in Table 3.11. The physical resource block number is $n_{PRB} = 10$ for the synchronization raster mapping to the synchronization block resource elements [9].

3.6 Frame Structure and Numerology

3GPP NR is designed to operate from 450 MHz to 100 GHz with a wide range of deployment scenarios, while supporting a variety of services. Given OFDMA was chosen as the multiple-access scheme for the downlink and uplink, it is not possible for a single numerology to satisfy the requirements of various use cases. Therefore, NR defines a family of OFDM numerologies for various frequency bands and deployment scenarios. The advantage of 3GPP NR relative to LTE is that it defines multiple numerologies which can be mixed and used simultaneously. A numerology is defined by a subcarrier spacing and a CP. The requirements for the OFDM subcarrier spacing is determined based on the carrier frequency, phase noise, delay spread, and Doppler spread. The use of smaller subcarrier spacing would result in either large EVM due to phase noise or more stringent requirements on the local oscillator. The small subcarrier spacing further leads to performance degradation in high Doppler scenarios. The required CP overhead and thus anticipated delay spread sets an

Table 3.12: Supported OFDM parameters in NR [12].

μ	Subcarrier Spacing $\Delta f = 2^{\mu} \times 15$ kHz	Cyclic Prefix Type	Supported for Data (PDSCH, PUSCH, etc.)	Supported for Synchronization Blocks (PSS, SSS, PBCH)	OFDM Useful Symbol Length μs	Cyclic Prefix Length μs	OFDM Symbol Length μs	N_{slot}^{symbol}	$N_{subframe}^{slot}$
0	15	Normal	Yes	Yes	66.67	4.69	71.35	14	1
1	30	Normal	Yes	Yes	33.33	2.34	35.68	14	2
2	60	Normal/ extended	Yes	No	16.67	1.17	17.84	14	4
3	120	Normal	Yes	Yes	8.33	0.57	8.92	14	8
4	240	Normal	No	Yes	4.17	0.29	4.46	14	16

upper limit for the subcarrier spacing. A large subcarrier spacing would result in unwanted overhead due to CP. The maximum FFT size of the OFDM modulation along with subcarrier spacing determines the channel bandwidth. Based on these observations, the subcarrier spacing should be as small as possible, while the system is still robust against phase noise and Doppler spread and supports the desired channel bandwidth. As shown in Table 3.12, in NR, the Primary and Secondary Synchronization Signals (PSS/SSS) and the Physical Broadcast Channel (PBCH), which are collectively known as SS block, will use 15/30 kHz SCS for sub-6 GHz and 120/240 kHz for above 6 GHz frequency bands. In the case of Physical Random-Access Channel (PRACH), the long preamble sequence utilizes 1.25/5 kHz SCS, in addition to the short preamble sequence using 15/30/60/120 kHz SCS. In other words, NR exploits a scalable OFDM subcarrier spacing (powers of 2) to support various frequency bands and deployment scenarios, where $\Delta f = 2^{\mu} \times 15\text{kHz}$ with $\mu = \{0, 1, 3, 4\}$ considered for PSS, SSS, and PBCH, and $\mu = \{0, 1, 2, 3\}$ designated for other physical channels. The normal CP is supported for all subcarrier spacing values, whereas extended CP is only supported for $\mu = 2$. From the network perspective, multiplexing of different numerologies over the same NR carrier bandwidth is possible in TDM and/or FDM manner in the downlink and uplink. From the UE perspective, multiplexing different numerologies is performed in TDM and/or FDM manner within or across a subframe. Regardless of the numerology used, the lengths of radio frame and subframe are always 10 and 1 ms, respectively. Different numerologies will then translate into the number of slots per subframe. The higher the subcarrier spacing, the more slots that can be accommodated per subframe [12]. Fig. 3.42 illustrates the NR frame, subframe, slot and mini-slot structure.

Different numerologies can be used in diverse deployment scenarios with their corresponding performance requirements. For example, the lower the subcarrier spacing, the larger the cell size, which will be suitable for the lower frequency deployments. At the same time, larger subcarrier spacing will allow shorter latency since the symbol duration will be shorter

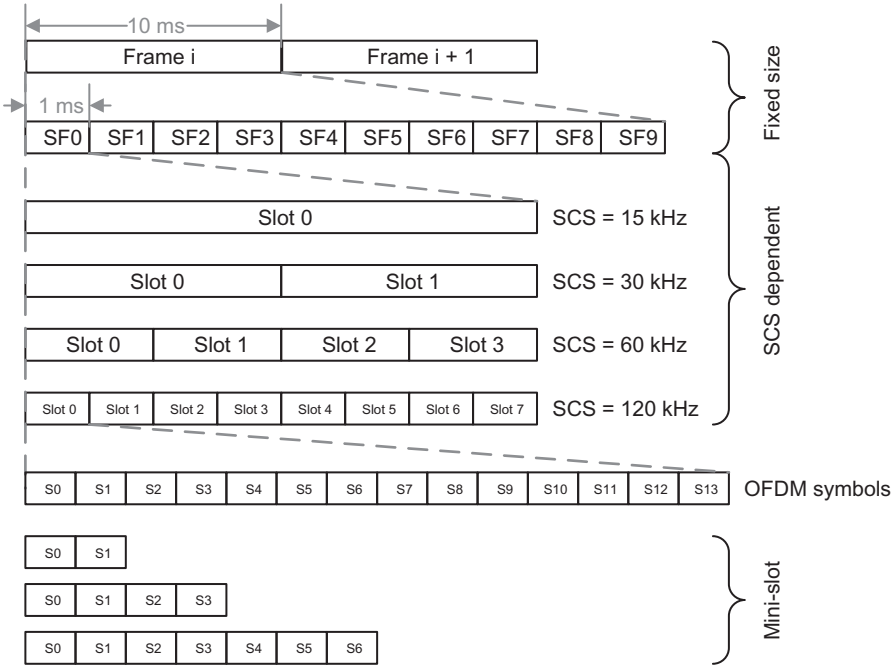


Figure 3.42
3GPP NR frame structure [12].

(see Fig. 3.42). Fig. 3.43 shows the relationship of the numerology, cell size, latency, and the carrier frequency [53].

The frame structure provides the basis for the timing of physical signal transmission. The timing scale is different for data, control, and synchronization physical channels. The sampling time in NR is defined as $T_c = 1/(\Delta f_{\max} N_{FFT})$ where $\Delta f_{\max} = 480\text{kHz}$ and $N_{FFT} = 4096$. In order to support multiple OFDM numerologies, the parameter μ and the corresponding CP for a BWP are signaled (configured) by the RRC parameters *DL-BWP-mu* and *DL-BWP-cp* for the downlink and *UL-BWP-mu* and *UL-BWP-cp* for the uplink [12,21]. The downlink and uplink transmissions are structured in the form of radio frames in the time domain with frame duration $T_{frame} = \left(\Delta f_{\max} N_{FFT}/100\right) T_c = 10$ ms, where each frame consists of ten subframes of $T_{subframe} = \left(\Delta f_{\max} N_f/1000\right) T_c = 1$ ms duration. The number of consecutive OFDM symbols in each subframe is defined as $N^{symbols}_{subframe}(\mu) = N^{symbols}_{slot} N^{slot}_{subframe}(\mu)$. In FDD mode, there is one set of frames in the uplink and one set of frames in the downlink on a carrier. As shown in Fig. 3.44, the ith uplink frame number starts at $T_{TA} = \left(N_{TA} + N_{TA\text{-}offset}\right) T_c$ before the start of the corresponding downlink frame at the UE in order to ensure uplink frame synchronization where $N_{TA\text{-}offset}$ depends on the frequency band of operation [18]. It can be seen that for the 15-kHz subcarrier spacing,

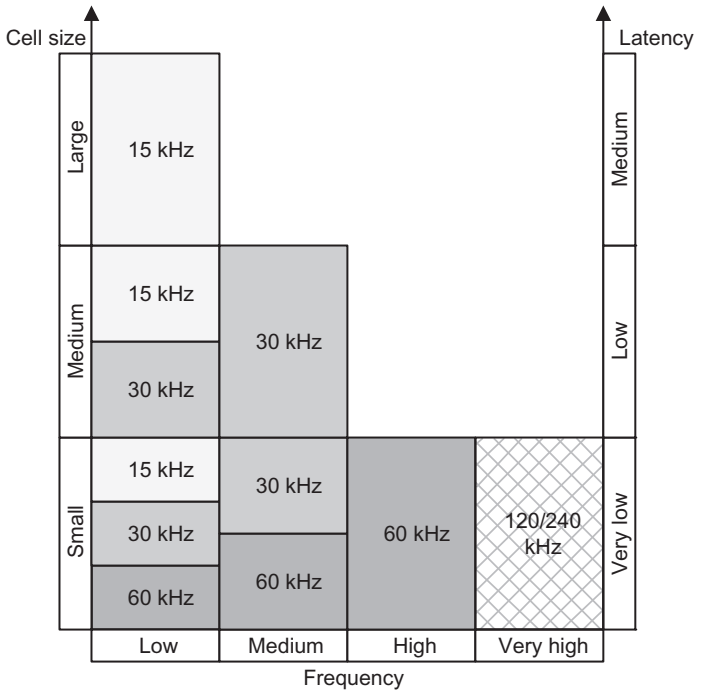
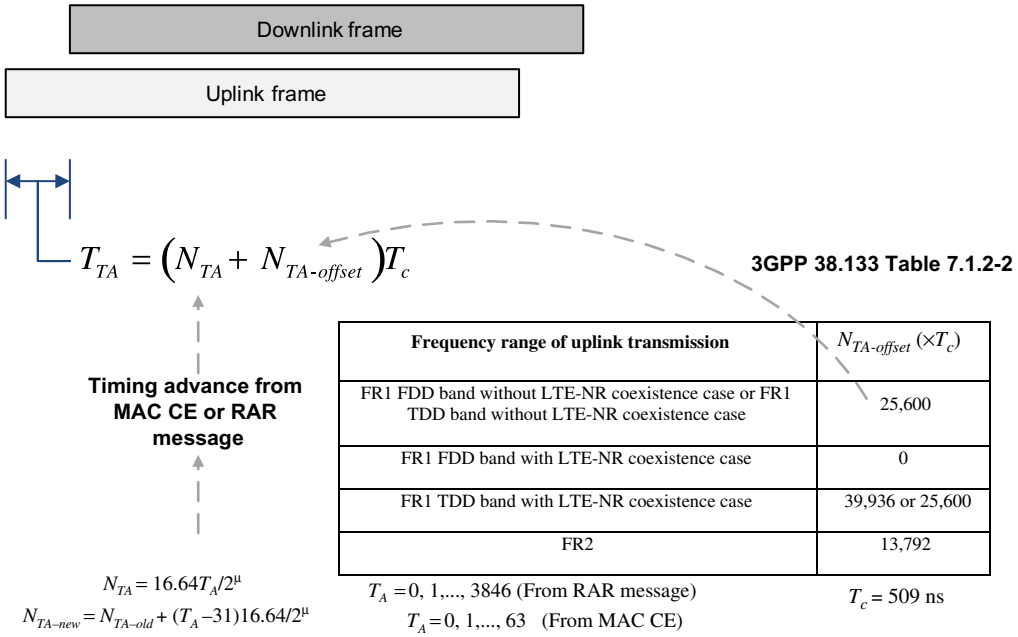


Figure 3.43
Relationship of numerology, carrier frequency, latency, and cell size [53].



Frequency range of uplink transmission	$N_{TA\text{-}offset}$ ($\times T_c$)
FR1 FDD band without LTE-NR coexistence case or FR1 TDD band without LTE-NR coexistence case	25,600
FR1 FDD band with LTE-NR coexistence case	0
FR1 TDD band with LTE-NR coexistence case	39,936 or 25,600
FR2	13,792

Figure 3.44
Illustration of uplink timing calculation in NR [12,39].

an NR slot has the same structure as the LTE subframe, which is important for supporting LTE/NR coexistence scenarios. In the case of co-located deployment, slot and frame structures may be aligned to simplify cell search and inter-frequency measurements [9,16,17,19]. Coordination of control signals and channels in time domain will also be possible to avoid interference between LTE and NR. Given that a slot is defined as a fixed number of OFDM symbols, a larger subcarrier spacing results in a shorter slot duration, which can be used to support low-latency applications.

3GPP NR further supports a more efficient approach to low-latency transmissions by allowing scheduling shorter slot sizes known as mini-slots (see Fig. 3.42). The mini-slot-based transmissions can also preempt an already ongoing slot-based transmission to another device, allowing immediate transmission of application data requiring very low latency. Mini-slots can be used for low-latency applications such as URLLC and operation in unlicensed bands, for example, to start transmission directly after a successful listen-before-talk procedure without waiting for the slot boundary. Mini-slots can consist of two, four, or seven OFDM symbols, where the first symbol includes (uplink or downlink) control information (see Fig. 3.42). For low-latency applications, the HARQ protocol can be configured either on a slot or a mini-slot basis. For the regular frame structure used by delay-tolerant applications, slot bundling as in LTE is also possible. Mini-slots may also be used for fast flexible scheduling of services (preemption of URLLC over eMBB). However, mini-slots are likely to be supported by few UE categories. A major difference between LTE and NR in terms of scheduling granularity is that LTE transmission time interval is fixed at 1 ms whereas NR transmission interval is a slot or a fraction of slot whose length is a function of the subcarrier spacing.

For a given subcarrier spacing parameter μ, the slots are numbered as $n_s = \left\{0, 1, \ldots, N_{subframe}^{slot} - 1\right\}$ in ascending order within a subframe. There are N_{slot}^{symbol} consecutive OFDM symbols in a slot where N_{slot}^{symbol} depends on the CP. The start of slot n_{slot} in a subframe is aligned in time with the start of OFDM symbol $n_{slot}N_{slot}^{symbol}$ in the same subframe [12]. In the TDD mode, the OFDM symbols in a slot can be classified as downlink, flexible, or uplink. Table 3.13 shows possible slot formats. In a downlink frame slot, the UE assumes that downlink transmissions only occur in downlink or flexible symbols, whereas in an uplink frame slot, the UE only transmits in uplink or flexible symbols.

In LTE TDD, there are a number of predefined patterns for uplink/downlink OFDM symbol allocation in a radio frame, while NR does not define any preset uplink/downlink patterns (see Fig. 3.45). A slot format indication (SFI) parameter informs the UE whether an OFDM symbol is downlink, uplink or flexible. The SFI can indicate link direction over one or many slots when configured through RRC. The SFI carries an index to Table 3.13 (pre-configured UE-specific slot configuration table) configured through RRC. The SFI can be either dynamically configured through a DCI or statically or semi-statically configured through RRC. The UE assumes there is no conflict between dynamic SFI and downlink control

Table 3.13: Slot formats for normal cyclic prefix ("D/U" denotes flexible downlink/uplink symbols) [12].

Slot Format	Symbol Number in a Slot													
	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL
1	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL
2	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U
3	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	D/U
4	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	D/U	D/U
5	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	D/U	D/U	D/U
6	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	D/U	D/U	D/U	D/U
7	DL	DL	DL	DL	DL	DL	DL	DL	DL	D/U	D/U	D/U	D/U	D/U
8	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	UL
9	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	UL	UL
10	D/U	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL
11	D/U	D/U	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL
12	D/U	D/U	D/U	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL
13	D/U	D/U	D/U	D/U	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL
14	D/U	D/U	D/U	D/U	D/U	UL	UL	UL	UL	UL	UL	UL	UL	UL
15	D/U	D/U	D/U	D/U	D/U	D/U	UL	UL	UL	UL	UL	UL	UL	UL
16	DL	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U
17	DL	DL	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U
18	DL	DL	DL	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U
19	DL	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	UL
20	DL	DL	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	UL
21	DL	DL	DL	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	UL
22	DL	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	UL	UL
23	DL	DL	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	UL	UL
24	DL	DL	DL	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	UL	UL
25	DL	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	UL	UL	UL
26	DL	DL	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	UL	UL	UL
27	DL	DL	DL	D/U	D/U	D/U	D/U	D/U	D/U	D/U	D/U	UL	UL	UL
28	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	D/U	UL
29	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	D/U	D/U	UL
30	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	D/U	D/U	D/U	UL
31	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	D/U	UL	UL

(*Continued*)

Table 3.13: (Continued)

Slot Format	Symbol Number in a Slot													
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
32	DL	DL	DL	DL	DL	DL	DL	DL	DL	DL	D/U	D/U	UL	UL
33	DL	DL	DL	DL	DL	DL	DL	DL	DL	D/U	D/U	D/U	UL	UL
34	DL	D/U	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL
35	DL	DL	D/U	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL
36	DL	DL	DL	D/U	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL
37	DL	D/U	D/U	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL
38	DL	DL	D/U	D/U	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL
39	DL	DL	DL	D/U	D/U	UL	UL	UL	UL	UL	UL	UL	UL	UL
40	DL	D/U	D/U	D/U	UL	UL	UL	UL	UL	UL	UL	UL	UL	UL
41	DL	DL	D/U	D/U	D/U	UL	UL	UL	UL	UL	UL	UL	UL	UL
42	DL	DL	DL	D/U	D/U	D/U	UL	UL	UL	UL	UL	UL	UL	UL
43	DL	DL	DL	DL	DL	DL	DL	DL	DL	D/U	D/U	D/U	D/U	UL
44	DL	DL	DL	DL	DL	DL	D/U	D/U	D/U	D/U	D/U	D/U	UL	UL
45	DL	DL	DL	DL	DL	DL	D/U	D/U	UL	UL	UL	UL	UL	UL
46	DL	DL	DL	DL	DL	D/U	UL	DL	DL	DL	DL	DL	D/U	UL
47	DL	DL	D/U	UL	UL	UL	UL	DL	DL	D/U	UL	UL	UL	UL
48	DL	D/U	UL	UL	UL	UL	UL	DL	D/U	UL	UL	UL	UL	UL
49	DL	DL	DL	DL	D/U	D/U	UL	DL	DL	DL	DL	D/U	D/U	UL
50	DL	DL	D/U	D/U	UL	UL	UL	DL	DL	D/U	D/U	UL	UL	UL
51	DL	D/U	D/U	UL	UL	UL	UL	DL	D/U	D/U	UL	UL	UL	UL
52	DL	D/U	D/U	D/U	D/U	D/U	UL	DL	D/U	D/U	D/U	D/U	D/U	UL
53	DL	DL	D/U	D/U	D/U	D/U	UL	DL	DL	D/U	D/U	D/U	D/U	UL
54	D/U	D/U	D/U	D/U	D/U	D/U	D/U	DL	DL	DL	DL	DL	DL	DL
55	DL	DL	D/U	D/U	D/U	UL	UL	UL	DL	DL	DL	DL	DL	DL
56–255	Reserved													

information (DCI) DL/UL assignments, thus when operating in NR TDD mode, one has to clearly define how available time slots are allocated to downlink and uplink transmissions. The NR defines those patterns in more flexible manner using the following parameters (see Fig. 3.46) [21]:

- *dl-UL-TransmissionPeriodicity*: Periodicity of the DL–UL pattern
- *nrofDownlinkSlots*: Number of consecutive full DL slots at the beginning of each DL–UL pattern

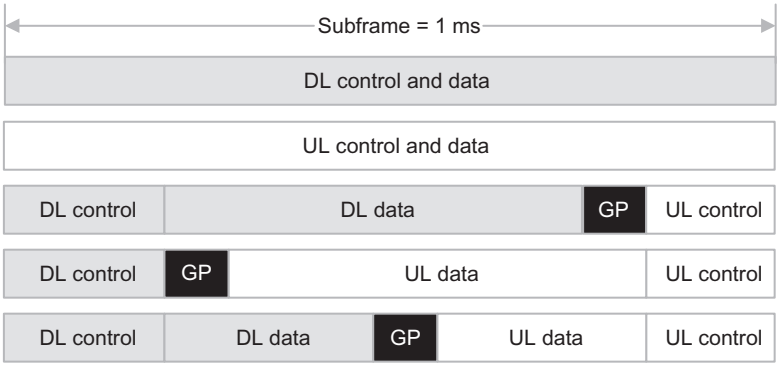


Figure 3.45
Different NR TDD-based subframe structures [53].

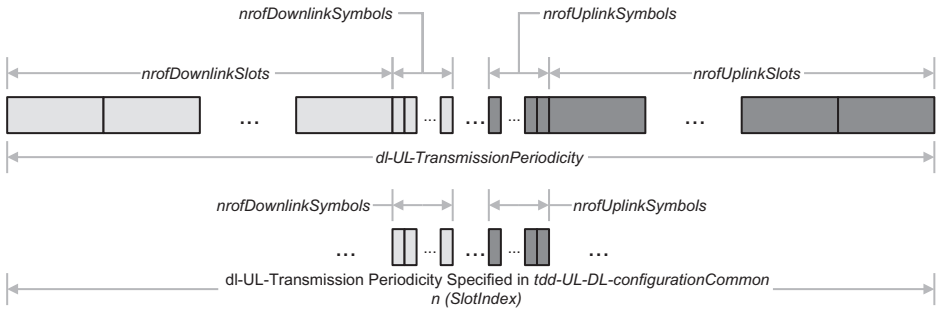


Figure 3.46
TDD UL/DL common and dedicated configurations [15,39].

- *nrofDownlinkSymbols*: Number of consecutive DL symbols in the beginning of the slot following the last full DL slot
- *nrofUplinkSlots*: Number of consecutive full UL slots at the end of each DL−UL pattern
- *nrofUplinkSymbols*: Number of consecutive UL symbols at the end of the slot preceding the first full UL slot

As shown in Table 3.13, a slot format includes a specific downlink, uplink, and flexible symbol configuration. In TDD mode, a slot can be all downlink, all uplink, or a combination of downlink and uplink segments. Data transmission can be scheduled to span one or multiple slots when slot aggregation is supported. For each serving cell, if a UE receives the RRC parameter *UL-DL-configuration-common*, it must set the slot format per slot over the number of slots indicated by this parameter. If the UE is additionally provided with RRC parameter *UL-DL-configuration-dedicated* for the slot format per slot over a number of slots, the latter parameter overrides only flexible symbols per slot over the number of slots indicated by *UL-DL-configuration-common* parameter. The UE determines the duration

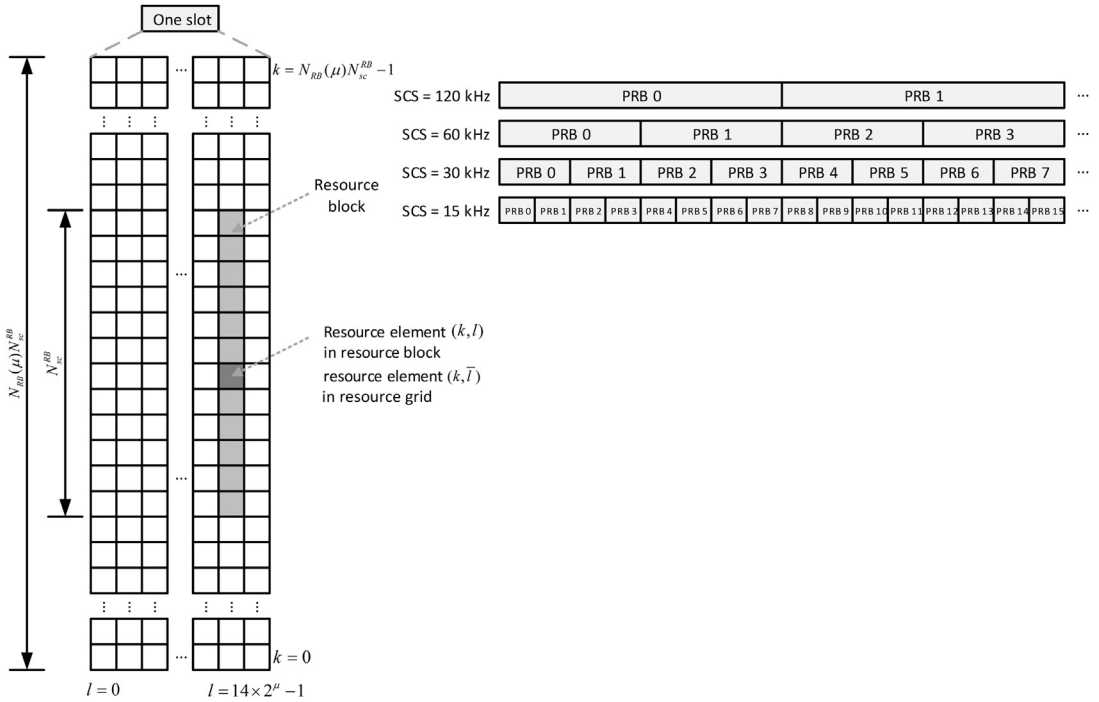


Figure 3.47
Illustration of NR resource grids and PRBs [12].

of each slot in the number of slots, in each configured BWP, based on the subcarrier spacing value provided by higher layer parameter *ref-scs*. The UE considers symbols in a slot indicated as downlink by higher layer parameter *UL-DL-configuration-common* or by higher layer parameter *UL-DL-configuration-dedicated* as available for receiving control/traffic. The UE further considers symbols in a slot as uplink indicated by higher layer parameter *UL-DL-configuration-common* or by higher layer parameter *UL-DL-configuration-dedicated* as available for transmission of control/traffic [14,15].

3.7 Time-Frequency Resources

3.7.1 Physical Resource Blocks

The basic scheduling unit in NR is a physical resource block (PRB) comprising 12 subcarriers in the frequency domain over one OFDM symbol. All subcarriers within a PRB have the same subcarrier spacing and CP length. When an NR system supports multiple numerologies, the corresponding PRBs are multiplexed in the time domain such that the boundaries of PRBs are aligned. For this purpose, multiple PRBs of the same bandwidth form a PRB grid, as illustrated in Fig. 3.47. A PRB grid formed by subcarriers spaced apart by

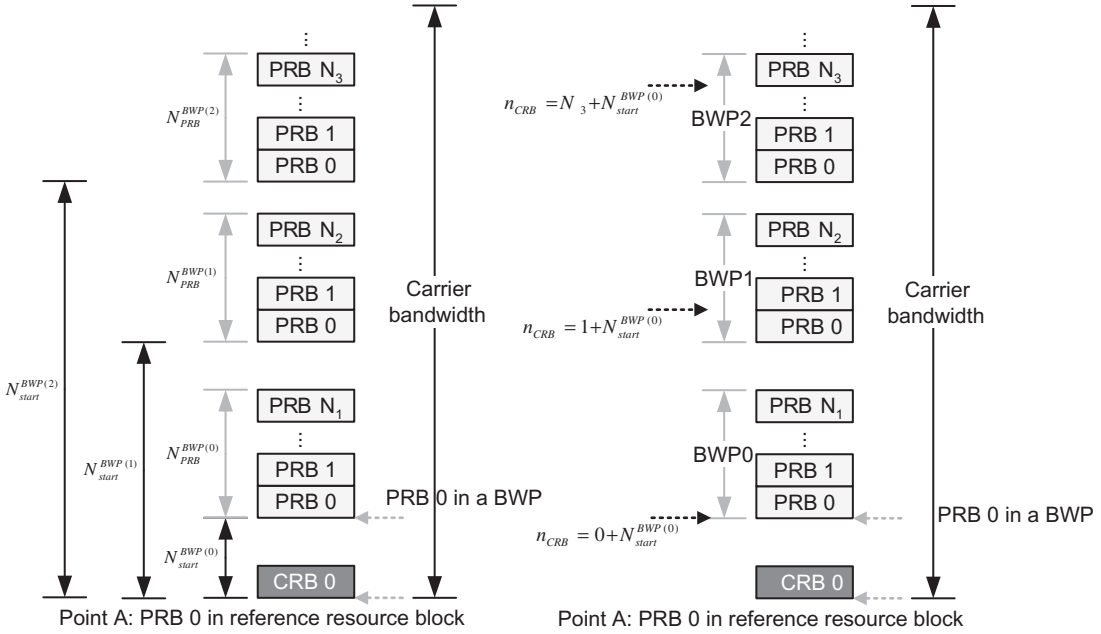


Figure 3.48
Mapping between n_{CRB} and n_{PRB} [12,39].

$\Delta f = 2^{\mu} \times 15$ kHz, where $\mu = 0, 1, \ldots, 4$ is a non-negative integer, is a superset of PRB grids with subcarrier spacing 15 kHz. For each numerology and carrier frequency, a resource grid of $N_{grid}(\mu)N_{sc}^{RB}$ subcarriers and $N_{symbol}(\mu)$ OFDM symbols is defined, starting at a CRB $N_{start}(\mu)$, whose value is signaled via RRC signaling. There is one set of resource grids per link direction (uplink or downlink). There is a single resource grid for a given antenna port p, numerology parameter μ and link direction [12,28].

An antenna port is a logical entity which is distinct from a physical antenna. Each antenna port is associated with a specific set of reference signals such that the channel over which a symbol is transmitted on that antenna port can be distinguished from the channel over which another symbol is conveyed on the same antenna port. Two antenna ports are said to be quasi-co-located, if the large-scale properties of the channel over which a symbol is conveyed on one antenna port can be inferred from the channel over which a symbol is conveyed on another antenna port. The large-scale properties include delay spread, Doppler spread, Doppler shift, average gain, average delay, and other spatial parameters [12]. In other words, a UE receiver can assume that the radio channels corresponding to two different antenna ports have the same large-scale properties (e.g., average delay spread, Doppler spread/shift, average delay, average gain, and spatial receive parameters), if the antenna ports are specified as being quasi-co-located. The UE can assume that two antenna ports are quasi-co-located with respect to certain channel properties either by NR specification or explicitly informed by the network via signaling.

Table 3.14: Minimum and maximum number of resource blocks/transmission bandwidths [12].

μ	min(N_{PRB})	max(N_{PRB})	Subcarrier Spacing (kHz)	Minimum Bandwidth (MHz)	Maximum Bandwidth (MHz)
0	24	275	15	4.32	49.5
1	24	275	30	8.64	99
2	24	275	60	17.28	198
3	24	275	120	34.56	396
4	24	138	240	69.12	397.44

Each element in the resource grid for antenna port p and numerology parameter μ is called a resource element and is uniquely identified by pair (k, l) where k is the index in the frequency domain (k is defined relative to point A such that $k = 0$ corresponds to the subcarrier centered around point A) and l refers to the symbol position in the time domain. A resource block is defined as $N_{sc}^{RB} = 12$ consecutive subcarriers in the frequency domain. Point A is a common reference point for resource block grids which is derived from the higher layer parameters [12].

As shown in Fig. 3.48, the resource blocks for each subcarrier spacing configuration are numbered from 0 to $N_{RB}(\mu)N_{sc}^{RB} - 1$ in upward direction in the frequency domain. Table 3.14 provides the minimum and maximum values of $N_{RB}(\mu)$ and their corresponding transmission bandwidths. The relationship between the CRB number n_{CRB} in the frequency domain and resource elements (k, l) for each subcarrier spacing configuration is given by $n_{CRB} = \lfloor k/N_{sc}^{RB} \rfloor$ where index k is defined relative to Point A such that $k = 0$ corresponds to the subcarrier centered around that point A. Physical resource blocks are defined within BWPs and are numbered from 0 to $N_{PRB}^{BWP(i)} - 1$. The relationship between the physical resource block n_{PRB} in the ith BWP and the CRB n_{CRB} is given by $n_{CRB} = n_{PRB} + N_{start}^{BWP(i)}$ where $N_{start}^{BWP(i)}$ is the CRB where BWP starts relative to CRB 0 as shown in Fig. 3.48. Similar to LTE, virtual resource blocks are defined within a BWP and are enumerated from 0 to $N_{VRB}^{BWP(i)} - 1$. The virtual resource blocks are resource blocks that are permuted across frequency dimension to take advantage of frequency diversity [12,28]. An interleaved mapping maps virtual resource blocks to physical resource blocks using an interleaver that spans the entire BWP and operates on pairs or quadruplets of resource blocks. A block interleaver with two rows is used, with pairs/quadruplets of resource blocks written in columns and read in rows. Whether to use pairs or quadruplets of resource blocks in the interleaving operation is configurable by higher layer signaling. The interleaved resource-block mapping in the frequency domain provides frequency diversity [13–15].

3.7.2 Bandwidth Part

3GPP NR supports very large operating bandwidths relative to the previous generations of 3GPP standards. Since the UEs in a cell may have different bandwidth capabilities, the use

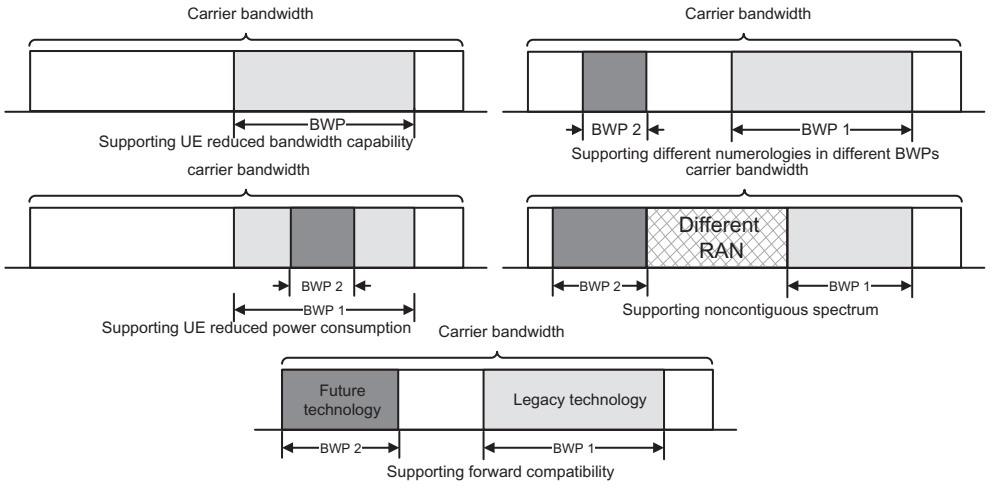


Figure 3.49
Bandwidth part use cases [42].

of wide bandwidth may cause more power consumption and may increase RF and baseband implementation complexity. Therefore, NR introduces the concept of BWP and allows the UEs with different bandwidth capabilities to operate in the cell with (configurable) smaller instantaneous bandwidth relative to the configured cell bandwidth, making NR more energy efficient solution despite its support of wideband channels (see Fig. 3.49). Alternatively, one may consider scheduling a UE such that it only transmits or receives within a certain frequency band. However, the difference of the latter approach with BWP is that the UE is not required to transmit or receive outside of the configured frequency band of the active BWP. The granularity of bandwidth allocation in NR is one PRB. For each serving cell, up to 4 downlink/uplink BWPs can be configured separately and independently for paired spectrum; nevertheless, only one BWP can be active at a given time and the UE is not expected to receive downlink/uplink physical signals/channels outside of an active BWP. For paired spectrum, a downlink BWP and an uplink BWP are jointly configured as a pair and up to four pairs can be configured. One can configure up to four BWPs on a supplemental uplink (SUL) carrier. Different use cases of the BWP are illustrated in Fig. 3.49 [28].

In other words, a BWP is a subset of contiguous CRBs for a given numerology (note that different numerologies may be used in different BWPs) on a given RF carrier. The starting position $N_{start}^{BWP(i)}$ and the number of resource blocks $N_{PRB}^{BWP(i)}$ in a BWP satisfy $N_{start}^{grid}(\mu) \leq N_{start}^{BWP(i)} < N_{start}^{grid}(\mu) + N_{size}^{grid}(\mu)$ and $N_{start}^{grid}(\mu) < N_{PRB}^{BWP(i)} + N_{start}^{BWP(i)} \leq N_{start}^{grid}(\mu) + N_{size}^{grid}(\mu)$, respectively. If a UE is configured with a SUL, then it can additionally be configured with up to four BWPs on the SUL with a single SUL BWP being active at a given time.

As we mentioned earlier, the transmit/receive bandwidth of a UE does not need to be as large as the bandwidth of the cell and it can be adaptively adjusted according to UE

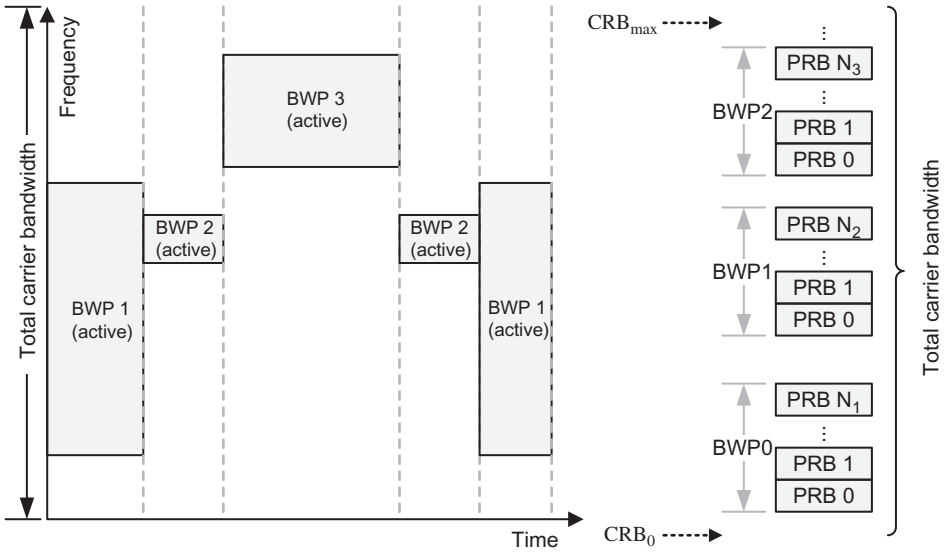


Figure 3.50
Illustration of the active bandwidth part adaptation concept [12,19,28].

operational conditions. With Bandwidth Adaptation (BA), a UE's bandwidth can be resized (e.g., reduced during a period of low activity for power saving); its location can be moved in the frequency domain (e.g., to increase scheduling flexibility); and the subcarrier spacing can be changed (e.g., to allow different services). A subset of the total cell bandwidth is called a BWP and the BA is achieved by configuring the UE with different BWP(s) and notifying the UE of the instantaneous active one. Fig. 3.50 shows a scenario where three different BWPs are configured: $BWP_1 = 50$ MHz and subcarrier spacing of 15 kHz; $BWP_2 = 10$ MHz and subcarrier spacing of 15 kHz; and $BWP_3 = 25$ MHz and subcarrier spacing of 60 kHz [19]. Note that at each time, only one BWP is active.

An initial active downlink BWP is defined by its location, number of contiguous PRBs, subcarrier spacing, and CP, for the control resource set corresponding to Type0-PDCCH common search space. For operation on the primary cell, a UE is provided by higher layer parameter *initial-UL-BWP* an initial uplink BWP to perform random-access procedure. If the UE is configured with a secondary carrier on the primary cell, it can also be configured with an initial BWP for random-access procedure on the secondary carrier [14,15]. A UE can be provided with a timer value by RRC parameter *BWP-InactivityTimer* for the primary cell. The UE subsequently starts the timer each time that it detects a DCI format 1_1 indicating an active downlink BWP, other than the default downlink BWP, for paired spectrum operation or each time the UE detects DCI format 1_1 or DCI format 0_1 indicating an active downlink or uplink BWP, other than the default downlink/uplink BWP, for unpaired spectrum operation. The UE increments the timer every 1 ms for sub-6 GHz carrier frequencies or every 0.5 ms for carrier frequencies above 6 GHz, if it does not detect a DCI format 1_1 for paired spectrum operation

or if it does not detect a DCI format 1_1 or DCI format 0_1 for unpaired spectrum operation. The timer expires when the value is equal to the *BWP-InactivityTimer*. Upon expiration of the timer, the UE switches from the active BWP to the default BWP [14,15].

In conjunction with an UL/DL carrier pair for an FDD band, or a bidirectional carrier for a TDD band, a UE may be configured with an additional SUL. The SUL differs from the aggregated uplink in the sense that the UE may be scheduled to transmit either on the SUL or on the uplink of the carrier being supplemented, but not on both at the same time. In the case of SUL, the UE is configured with two uplink carriers in conjunction of one downlink carrier of the same cell, and uplink transmissions on those carriers are controlled by the network to avoid colliding uplink control and traffic channels in time domain. The colliding transmissions on uplink traffic channel are avoided through scheduling while overlapping transmissions on uplink control channel are avoided via limiting configuration of uplink control channel on only one of the two uplink carriers. In addition, initial access is supported on each of the uplink carriers. To improve uplink coverage for high-frequency scenarios, a low-frequency SUL carrier can be configured. In NR, the UE can take advantage of the bandwidth adaptation feature to save power while satisfying the requirements of various services/applications. The network can configure up to four BWPs for each UE, and dynamically send change indications to the UEs as required.

The BWP switching for a serving cell is used to activate an inactive BWP or to deactivate an active BWP at any given time, as shown in Fig. 3.51. The BWP switching is controlled by the PDCCH indicating a downlink assignment or an uplink grant, by the *bandwidthPartInactivityTimer*, or by the MAC entity itself upon initiation of random-access procedure. Upon addition of a Special Cell (SpCell)[22] or activation of an SCell, one BWP is initially active without receiving PDCCH indicating a downlink assignment or an uplink grant. The active BWP for a serving cell is indicated by either RRC or PDCCH. For paired spectrum, a downlink BWP is paired with an uplink BWP, and BWP switching applies to both uplink and downlink [20]. The BWP switching options are illustrated and compared in Fig. 3.52.

3.7.3 Resource Allocation

One of the main design objectives for signaling the resource allocation information, in the form of a set of resource blocks in each slot, to the active UEs in the cell is to find a balanced tradeoff between flexibility and signaling overhead. Indications of localized/distributed resource allocations to different UEs are transmitted via PDCCH. The resource allocation field in PDCCH is interpreted by the UE depending on the PDCCH DCI format.

[22] In the context of dual connectivity, the Special Cell refers to the primary cell of the MCG or the primary SCell of the SCG depending on whether the MAC entity is associated with the MCG or the SCG. Otherwise, the Special Cell refers to the PCell. A Special Cell supports PUCCH transmission and contention-based random-access procedure and is always activated [20].

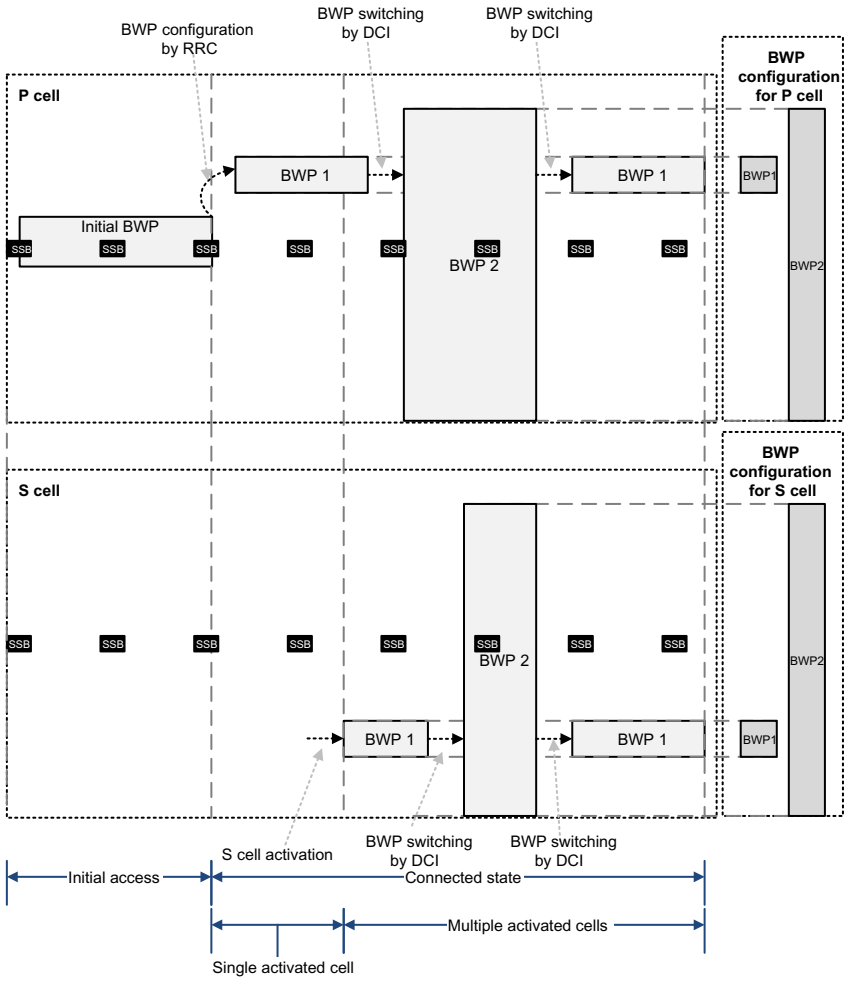


Figure 3.51
Illustration of BWP adaptation, activation, and switching [42].

The resource allocation in NR is defined both in time domain and frequency domain. Unlike LTE where the resource allocation in time domain was determined based on a fixed/predefined rule, in NR the resource allocation in time domain is more flexible, whereas the resource allocation in frequency domain is relatively similar to that of LTE [30]. Resource allocation type specifies the way in which the scheduler allocates physical resource blocks in frequency domain to each user for transmission in the downlink or uplink.

3.7.3.1 Resource Allocation in Time Domain

In the downlink, when the UE is scheduled to receive PDSCH via a DCI, that is, the *time domain resource assignment* field of DCI provides a row index to an allocation table, where the indexed row defines the slot offset K_0, the start and length indicator *SLIV* and the

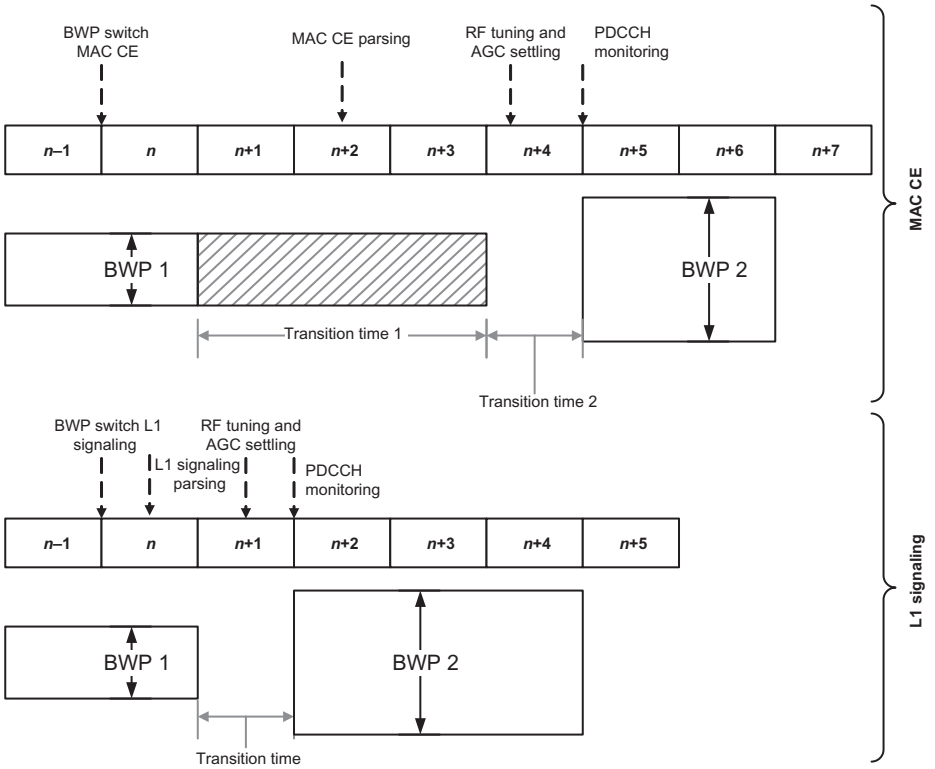


Figure 3.52
Comparison of BWP switching options [39].

PDSCH mapping type. As shown in Fig. 3.53, the slot allocated for PDSCH transmission is determined by parameter K_0 of the indexed row $n + K_0$ where n is the slot with the scheduling DCI, K_0 is based on the numerology of PDSCH. The starting symbol S relative to (the start of the slot), and the number of consecutive symbols L counting from the symbol S allocated for the PDSCH are determined from the start and length indicator $SLIV$ [14,15]. The slot allocated for the PDSCH is defined as $\left\lfloor \left(2^{\mu_{PDSCH}}/2^{\mu_{PDCCH}}\right)n \right\rfloor + K_0$, where μ_{PDSCH} and μ_{PDCCH} are the subcarrier spacing configurations for PDSCH and PDCCH, respectively. If $(L-1) \leq 7$, then $SLIV = 14(L-1) + S$; otherwise $SLIV = 14(14 - L + 1) + (14 - S - 1)$, where $0 < L \leq 14 - S$ and the PDSCH mapping type is set to Type A or Type B [15]. The permissible S and L combinations corresponding to PDSCH allocations are shown in Table 3.15.

The PDSCH mapping type is related to the relative location of the demodulation reference signal (DM-RS) and the slot boundary as well as the size of the data. The mapping type A is used when the first DM-RS is located in the second or third OFDM symbol of

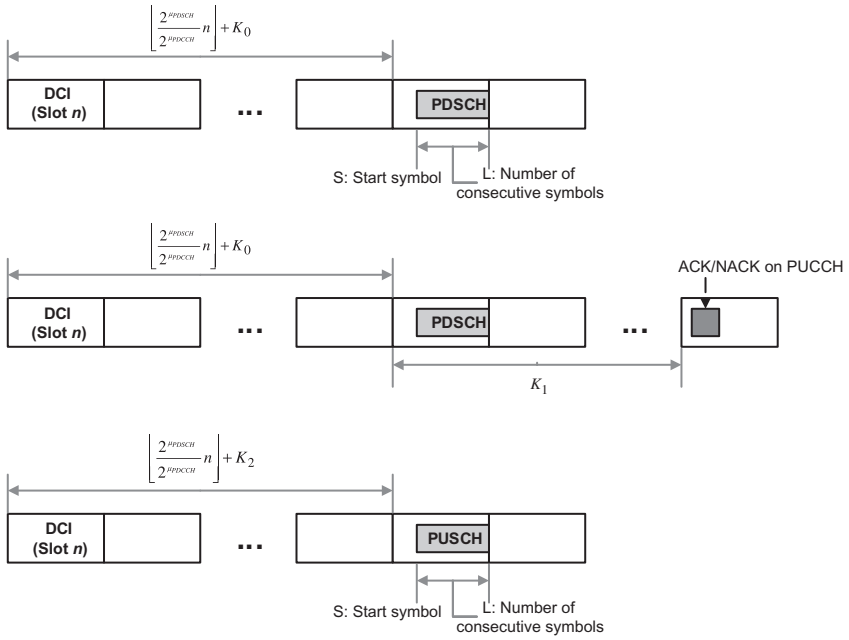


Figure 3.53
Illustration of downlink/uplink time-domain resource allocation [12,15,39].

Table 3.15: Permissible *S* and *L* values [15].

Mapping Type		Normal Cyclic Prefix			Extended Cyclic Prefix		
		S	***L***	***S*** + ***L***	***S***	***L***	***S*** + ***L***
PDSCH	Type A	{0,1,2,3}	{3,. . .,14}	{3,. . .,14}	{0,1,2,3}	{3,. . .,12}	{3,. . .,12}
	Type B	{0,. . .,12}	{2,4,7}	{2,. . .,14}	{0,. . .,10}	{2,4,6}	{2,. . .,12}
PUSCH	Type A	0	{4,. . .,14}	{4,. . .,14}	0	{4,. . .,12}	{4,. . .,12}
	Type B	{0,. . .,13}	{1,. . .,14}	{1,. . .,14}	{0,. . .,12}	{1,. . .,12}	{1,. . .,12}

the slot following a CORESET (i.e., a control region) at the beginning of a slot. The DM-RS is mapped relative to the start of the slot boundary regardless of the start of data transmission in the slot. This mapping type is primarily intended for the cases where the data occupies most of the slot. The mapping type B is used when the first DM-RS is located in the first symbol of the data allocation, that is, the DM-RS location is not given relative to the slot boundary, rather relative to where the data is located. This mapping is intended for transmissions over a small fraction of the slot to support very low latency and other transmissions that cannot wait until a slot boundary starts regardless of the transmission duration.

When the UE is configured with $pdsch{-}AggregationFactor > 1$, the same symbol allocation is applied across the *pdsch—AggregationFactor* consecutive slots that have not been defined as uplink by the SFI. Fig. 3.53 illustrates time-domain and frequency-domain resource allocation procedure for PDSCH and demonstrates how the above parameters are used to locate the user allocation.

In the uplink, when a UE is scheduled to transmit a transport block on PUSCH by a DCI with or without CSI report(s), the *time domain resource assignment* field of the DCI provides a row index to a table defined in [15], where the indexed row defines the slot offset K_2, the start and length indicator *SLIV*, or directly by the start symbol S and the allocation length L, and the mapping type to be used in PUSCH transmission as shown in Fig. 3.53. The slot where the UE transmits the PUSCH is determined by parameter K_2 as $\left\lfloor \left(2^{\mu_{PUSCH}}/2^{\mu_{PDCCH}}\right)n \right\rfloor + K_2$ where n is the slot with the scheduling DCI, K_2 is based on the numerology of PUSCH, μ_{PUSCH} and μ_{PDCCH} are the subcarrier spacing configurations for PUSCH and PDCCH, respectively. The starting symbol S relative to the start of the slot, and the number of consecutive symbols L counting from the symbol S allocated for the PUSCH are determined from the start and length indicator *SLIV* of the indexed row as follows. If $(L-1) \leq 7$ then $SLIV = 14(L-1) + S$; otherwise $SLIV = 14(14 - L + 1) + (14 - S - 1)$, where $0 < L \leq 14 - S$ and the PUSCH mapping type is set to Type A or Type B [15]. The permissible S and L combinations corresponding to PDSCH allocations are shown in Table 3.15. When the UE is configured with $pusch - AggregationFactor > 1$, the same symbol allocation is applied across the $pusch - AggregationFactor$ consecutive slots and the PUSCH is limited to a single transmission layer. The UE repeats the transport block across $pusch - AggregationFactor$ consecutive slots applying the same symbol allocation in each slot [15].

3.7.3.2 Resource Allocation in Frequency Domain

A UE determines the frequency-domain resources on which it transmits or receives data by examining the resource-block allocation and BWP indicator fields in a DCI. The resource allocation fields determine the resources blocks in the active BWP on which data is transmitted. The gNB can signal the allocated resources to a UE using resource allocation type 0 or type 1, which are conceptually similar to LTE resource allocation type 0 and type 2 with the difference that in LTE, the resource allocation signales the allocations across the carrier bandwidth, whereas in NR, the indication is relevant only for the active BWP. Resource allocation type 0 is a bitmap-based allocation scheme, indicating the set of resource blocks that the UE is supposed to receive in the downlink transmission where the size of the bitmap is equal to the number of resource blocks in the BWP. This would allow an arbitrary combination of resource blocks to be scheduled for the UE at the expense of a large control/signaling overhead and some downlink coverage issues for larger BWP sizes due to limited capacity of a single OFDM symbol. Consequently, there

Table 3.16: Nominal resource block group size P [15].

Bandwidth Part Size $N_{PRB}^{BWP(i)}$	Configuration 1P	Configuration 2P
1–36	2	4
37–72	4	8
73–144	8	16
145–275	16	16

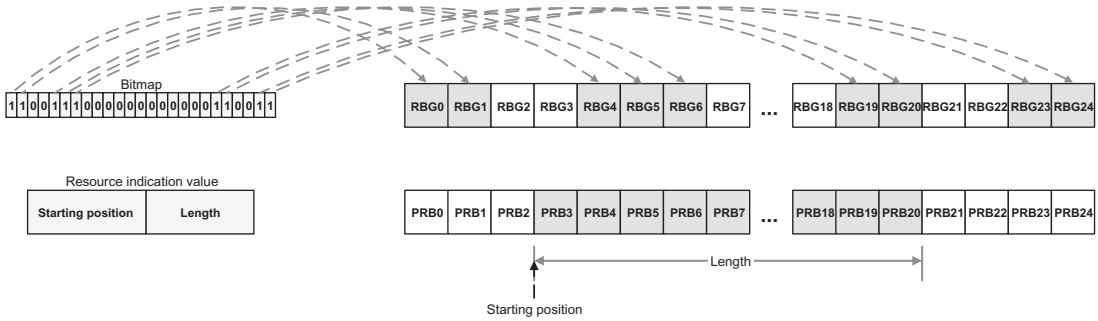


Figure 3.54
Illustration of frequency-domain type 0 and type 1 resource allocations [15,39].

is a need to reduce the bitmap size while maintaining the allocation flexibility. This can be achieved by addressing a group of contiguous resource blocks as opposed to individual PRBs. The size of the resource block group (RBG) is determined by the size of the BWP (see Table 3.16). Resource allocation type 1 indicates the allocated resources to the UE by means of a starting position and the length of the resource block allocation, thus only supporting contiguous allocations in frequency domain. In order to further reduce the signaling overhead, resource allocation type 1 combines the starting position and the length of resource allocation values into a single value referred to as resource indication value [15]. The two resource allocation types are illustrated in Fig. 3.54 (example).

The resource allocation scheme is configured using a bit in the DCI. Both resource allocation types refer to virtual resource blocks. For resource allocation type 0, a non-interleaved mapping from virtual to physical resource blocks is used, thus the virtual resource blocks are directly mapped to the corresponding physical resource blocks. For resource allocation type 1, both interleaved and non-interleaved mapping is supported. The VRB-to-PRB mapping bit, if present, indicates whether the allocation is based on interleaved or non-interleaved mapping.

In downlink resource allocation type 0, the resource block assignment information includes a bitmap representing the RBGs that are allocated to the scheduled UE where RBG is a set of consecutive physical resource blocks defined by a higher layer parameter and the size of the carrier BWP. The total number of RBGs N_{RBG} defined for a downlink carrier BWP of

size $N_{PRB}^{BWP(i)}$ is $N_{RBG} = \left\lceil \left[N_{PRB}^{BWP(i)} + \left(N_{PRB}^{BWP(i)} \mathrm{mod} P \right) \right] / P \right\rceil$ where the size of the first RBG is $RBG_0 = P - N_{start}^{BWP(i)} \mathrm{mod} P$, the size of last RBG is $RBG_{last} = \left(N_{start}^{BWP(i)} + N_{PRB}^{BWP(i)} \right) \mathrm{mod} P$, if $\left(N_{start}^{BWP(i)} + N_{PRB}^{BWP(i)} \right) \mathrm{mod} P > 0$ and P otherwise. The size of all other RBGs is P [15]. The nominal values of P are given in Table 3.16. The bitmap is of size N_{RBG} bits with one bit per RBG such that each RBG is directly addressable. The RBGs is indexed in the order of increasing frequency and starting at the lowest frequency of the carrier BWP. An RBG is allocated to the UE, if the corresponding bit value in the bitmap is set to one [15].

In downlink resource allocation type 1, the resource block assignment information indicates a set of contiguously allocated non-interleaved or interleaved virtual resource blocks within the active BWP of size $N_{PRB}^{BWP(i)}$ except for the case when DCI format 1_0 is decoded in any common search space in which case the size of CORESET 0 is used. The type 1 resource allocation field for the downlink consists of a Resource Indication Value (RIV) corresponding to a starting virtual resource block RB_{start} and a length in terms of contiguously allocated resource blocks L_{RB}. In that case, if $(L_{RB} - 1) \leq \left\lfloor N_{PRB}^{BWP(i)} / 2 \right\rfloor$ then $\mathrm{RIV} = N_{PRB}^{BWP(i)} (L_{RBs} - 1) + \mathrm{RB}_{start}$; otherwise $\mathrm{RIV} = N_{PRB}^{BWP(i)} (N_{PRB}^{BWP(i)} - L_{RB} + 1) + (N_{PRB}^{BWP(i)} - 1 - \mathrm{RB}_{start})$, where $1 \leq L_{RB} < N_{PRB}^{BWP(i)} - RB_{start}$ [15]. However, when the DCI size for DCI format 1_0 in UE-specific search space is derived from the size of CORESET 0 and is applied to another active BWP with the size of N_{BWP}^{active}, a downlink type 1 resource block assignment field consists of a resource indication value corresponding to a starting resource block $\mathrm{RB}_{start} = 0, K, 2K, \ldots, (N_{BWP}^{initial} - 1)K$ and a length in terms of virtually contiguously allocated resource blocks $L_{RBs} = K, 2K, \ldots, N_{BWP}^{initial} K$.

In the uplink, the UE determines the resource block assignment in frequency domain using the resource allocation field of DCI except for Msg.3 PUSCH initial transmission [15]. Two uplink resource allocations type 0 and type 1 are defined where resource allocation type 0 is used for PUSCH transmission when transform precoding is disabled. The uplink resource allocation type 1 is used for PUSCH transmission regardless of whether transform precoding is enabled or disabled. The UE can assume that when the scheduling PDCCH is received with DCI format 0_0, then uplink resource allocation type 1 is used. If a BWP indicator field is not configured in the scheduling DCI, the RB indexing for uplink type 0 and type 1 resource allocation is determined within the UE's active BWP. If a BWP indicator field is configured in the scheduling DCI, then the RB indexing for uplink type 0 and type 1 resource allocation is determined within the UE's BWP indicated by BWP indicator field value in the DCI. Upon detection of PDCCH intended for the UE, the UE must first determine the uplink BWP and then the resource allocation within the BWP [15].

The uplink resource allocation type 0 is similar to the downlink counterpart, where the resource block assignment information includes a bitmap indicating the RBGs that are

allocated to the scheduled UE. The size of RBGs is given in Table 3.16. The uplink resource allocation type 1 is also similar to the downlink part described earlier, where the resource block assignment information informs the UE of a set of contiguously allocated noninterleaved virtual resource blocks within the active carrier BWP of size N_{PRB}^{BWP} except for the case when DCI format 0_0 is decoded in any common search space in which case the size of the initial BWP $N_{PRB}^{BWP(0)}$ is used [15].

3.7.3.3 Physical Resource Block Bundling

A UE cannot make any assumption about correlation of reference signals between different PDSCH scheduling occasions in the time domain. This is necessary to allow more flexibility in precoder-based beamforming and spatial signal processing. In the frequency domain, however, the UE can assume that there is correlation between reference signals within a precoding resource block group (PRG). Over the frequency span of one PRG, the UE may assume that the downlink precoder remains the same, and exploit this in the channel estimation process. The correlation assumption does hold between PRGs. It can be concluded that there is a tradeoff between the precoding flexibility and the channel estimation performance, that is, a large PRG size can improve the channel estimation accuracy at the cost of precoding flexibility. Therefore, the gNB may indicate the PRG size to the device where the possible PRG sizes are two, four or the total scheduled bandwidth in terms of PRBs. It is also possible to dynamically indicate the PRG size through the DCI. In addition, the UE can be configured to assume that the PRG size is equal to the scheduled bandwidth when the scheduled bandwidth is larger than half of the active BWP.

A UE may assume that the precoding granularity is $P'_{BWP(i)}$ consecutive resource blocks in the frequency domain, where $P'_{BWP(i)}$ values are taken from the limited set of {2, 4, wideband}. If $P'_{BWP(i)}$ is set to "wideband", the UE can expect to be scheduled with contiguous PRBs with the PRG and the use of the same precoding across the allocated resources by the gNB. If $P'_{BWP(i)}$ is set to 2 or 4, the ith BWP is partitioned into $P'_{BWP(i)}$ consecutive PRBs to form the PRGs. In practice, the number of consecutive PRBs in each PRG can be one or more. The first PRG size is given by $P'_{BWP(i)} - N_{start}^{BWP(i)} \bmod P'_{BWP(i)}$ and the last PRG size given by $(N_{start}^{BWP(i)} + N_{PRB}^{BWP(i)}) \bmod P'_{BWP(i)}$, if $(N_{start}^{BWP(i)} + N_{PRB}^{BWP(i)}) \bmod P'_{BWP(i)} \neq 0$; otherwise, the last PRG size is $P'_{BWP(i)}$ [15]. The UE may assume the same precoding is applied for any downlink contiguous allocation of PRBs within a PRG. If a UE is scheduled a PDSCH with DCI format 1_0, it can assume that $P'_{BWP(i)}$ is equal to 2 PRBs [15].

3.7.4 Resource Allocation for Grant-Free/Semi-persistent Scheduling

In grant-free uplink transmissions, the UEs can transmit within a set of predetermined resource blocks without any explicit scheduling grants from the base station, resulting in lower control/signaling overhead and latency. However, uplink transmissions require the

UEs to transmit within a given set of resources that are pre-allocated by the base station with a certain periodicity. To avoid resource utilization inefficiency, multiple UEs might be allowed to share the same resources and to operate in a contention-based manner. Hence, collisions are inevitable, affecting connection reliability and latency, which is worsen as the UE traffic increases and retransmissions become necessary. In other words, grant-free transmission may not be scalable as the UE density increases in a network.

Fig. 3.55 shows a comparison between the contention-based (grant-free) and grant-based performance, where we observe that the efficiency of contention-based transmission degrades with increasing packet size and traffic load, while efficiency of grant-based transmission improves by increasing packet size and is stable over different load factors. The markings in the figure indicate that for packet sizes of 20, 30, and 40 bytes, there is a loading factor threshold below which contention-based transmission is more efficient than grant-based transmission. When the packet size is sufficiently small, for example, 10 bytes, contention-based transmission is always optimal within certain load factor range. Based on the analysis in [43], keep-alive messages of mobile Internet traffic are better suited to use

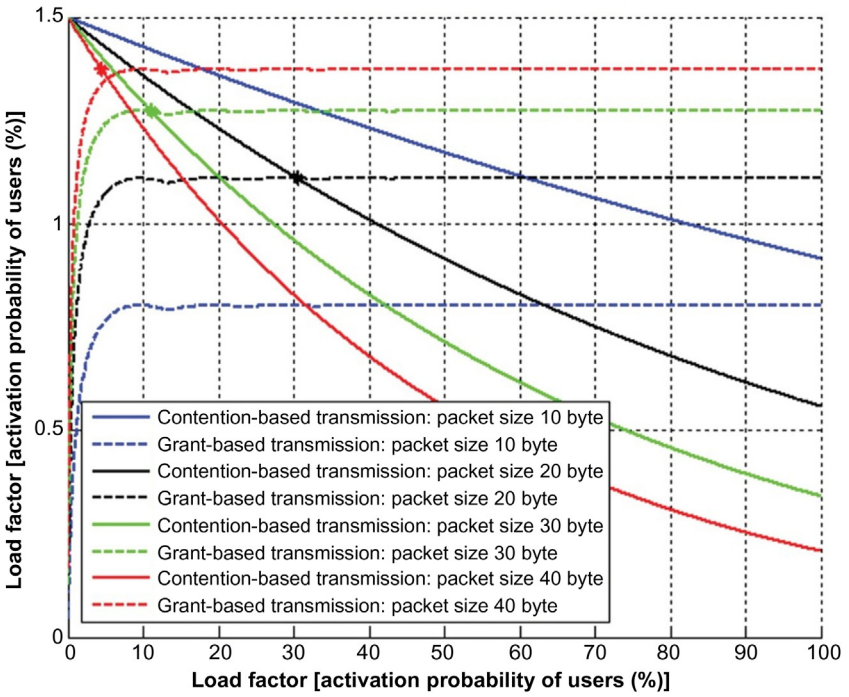


Figure 3.55
Comparison of grant-based and grant-free uplink transmissions [43].

contention-based transmission, while the rest of the mobile Internet packet types (e.g., video or voice) need to be transmitted by grant-based transmission, if not taking into account the latency constraints. This is because the former is of relatively small packet size and small loading, while packet size of the latter is large to be transmitted by contention (Fig. 3.55).

The semi-persistent scheduling (SPS)−based resource allocation refers to a transmission mode in which the serving base station allocates at least a part of resources and transport formats to the UE semi-statically over a certain time interval consisting of a number of TTIs. In the downlink, the semi-persistent scheduling is configured by RRC per serving cell and per BWP. Multiple configurations can be active simultaneously on different serving cells. Activation and deactivation of the downlink semi-persistent scheduling transmissions are independent among the serving cells [20].

In downlink semi-persistent scheduling, a downlink assignment is provided by PDCCH, and stored or cleared based on layer-1 signaling indicating semi-persistent scheduling activation or deactivation. A configured scheduling RNTI (CS-RNTI), which is similar to SPS C-RNTI in LTE, is configured by RRC for activation, deactivation, and retransmission. Furthermore, the number of configured HARQ processes and the periodicity of semi-persistent scheduling are signaled by RRC. When SPS resources are released by upper layers, all the corresponding configurations are subsequently released. In other words, the gNB can allocate downlink resources for the initial HARQ transmissions to UEs. The RRC defines the periodicity of the configured downlink assignments while PDCCH addressed with CS-RNTI can either signal and activate the configured downlink assignment, or deactivate it, that is, a PDCCH scrambled with CS-RNTI indicates that the downlink assignment can be implicitly reused according to the periodicity defined by RRC, until deactivated. After a downlink assignment is configured for semi-persistent scheduling, the MAC entity considers that the *N*th downlink assignment will occur in a slot whose number meets the following criteria:

$$\left(N_{frame}^{slot} SFN + n_{slot}\right) = \left[\left(N_{frame}^{slot} SFN_{start-time} + T_{start-time}^{slot}\right) + NT_{SPS} N_{frame}^{slot}/10\right] \bmod \left(1024 N_{frame}^{slot}\right)$$

where N_{frame}^{slot} denotes the number of slots per frame n_{slot} is the slot number in the frame, T_{SPS} denotes the SPS period, $SFN_{start-time}$ and $T_{start-time}^{slot}$ represent the SFN and slot of the first transmission of PDSCH where the configured downlink assignment was initialized, respectively [20].

In the uplink direction, there are two types of transmission without dynamic grant known as configured grant Type 1, where an uplink grant is provided by RRC, and stored as configured uplink grant; and configured grant Type 2, where an uplink grant is provided by PDCCH and stored or cleared as configured uplink grant based on physical layer signaling

indicating configured grant activation or deactivation. Both Type 1 and Type 2 grants are configured by RRC per serving cell and per BWP. Multiple configurations can be active simultaneously on different serving cells. For Type 2 grant, activation and deactivation are independent among the serving cells. When the configured grant Type 1 is used, the RRC configures the following parameters: a CS-RNTI for retransmission; periodicity of the configured grant Type 1; the offset of a resource with respect to $SFN = 0$ in time domain; time-domain parameters which include the start symbol and the length of the assignment; as well as the number of HARQ processes [20]. Alternatively, when the configured grant Type 2 is going to be used, the RRC configures the following parameters: a CS-RNTI for activation, deactivation, and retransmission; the periodicity of the configured grant Type 2; and the number of HARQ processes. Once a configured grant Type 1 semi-persistent allocation is set up in a serving cell by upper layers, the corresponding MAC entity stores the uplink grant provided by upper layers and initializes the configured uplink grant to start in the symbol according to the provided parameters. After an uplink grant is set up for a configured grant Type 1 uplink transmission, the MAC entity considers the Nth uplink transmit opportunity to occur in the symbol number which satisfies the following equation [20]:

$$\left[\left(N_{frame}^{slot}N_{slot}^{symbol}SFN\right) + \left(n_{slot}N_{slot}^{symbol}\right) + m_s\right] = \left(t_{offset}N_{slot}^{symbol} + S + NT_{SPS}\right)\mathrm{mod}\left(1024N_{frame}^{slot}N_{slot}^{symbol}\right), \quad \forall N \geq 0$$

where N_{slot}^{symbol}, m_s, and t_{offset} denote the number of symbols per slot, symbol number in the slot, and the time domain offset, respectively. Similarly, subsequent to an uplink grant set up for a configured grant Type 2, the MAC entity considers the time-domain location of the Nth uplink grant-free transmission at the symbol for which the following criterion is satisfied [20]:

$$N_{frame}^{slot}N_{slot}^{symbol}SFN + n_{slot}N_{slot}^{symbol} + m_s = \left(SFN_{start-time}N_{frame}^{slot}N_{slot}^{symbol} + T_{start-time}^{slot}N_{slot}^{symbol} + T_{start-time}^{symbol} + NT_{SPS}\right)\mathrm{mod}\left(1024N_{frame}^{slot}N_{slot}^{symbol}\right), \quad \forall N \geq 0$$

where $SFN_{start-time}, T_{start-time}^{slot}$, and $T_{start-time}^{symbol}$ represent the SFN, slot, and symbol where the first transmission of PUSCH with the configured uplink grant was initialized, respectively. When a configured grant is released by upper layers, all corresponding configurations are cleared. Retransmissions except for repetition of the configured grants use uplink grants with CS-RNTI.

In summary, in the downlink, the gNB can allocate downlink resources for the initial HARQ transmissions to UEs with semi-persistent scheduling. The RRC signaling defines the periodicity of the configured downlink assignments while PDCCH addressed with

CS-RNTI can either signal and activate the configured downlink assignment, or deactivate it, that is, a PDCCH scrambled with CS-RNTI indicates that the downlink assignment can be implicitly reused according to the periodicity defined by RRC, until deactivated. In the uplink, the gNB can allocate uplink resources for the initial HARQ transmissions to UEs with configured uplink grants. Two types of configured uplink grants are defined: Type 1, where RRC directly provides the configured uplink grant (including the periodicity) and Type 2, where RRC defines the periodicity of the configured uplink grant while PDCCH addressed to CS-RNTI can either signal and activate the configured uplink grant, or deactivate it, that is, a PDCCH addressed with CS-RNTI indicates that the uplink grant can be implicitly reused according to the periodicity defined by RRC, until deactivated [19].

When carrier aggregation is configured, one configured uplink grant can be signaled per serving cell. Thus each serving cell can have one configured uplink grant active at any time. Similarly, when bandwidth adaptation is configured, one configured uplink grant can be signaled per BWP. A configured uplink grant for one serving cell can either be of Type 1 or Type 2. For Type 2, activation and deactivation of configured uplink grants are independent among the serving cells. When SUL is configured, a configured uplink grant can only be signaled for one of the two uplink carriers of the cell.

References

ITU-R Specifications[23]

[1] Report ITU-R P.2406-0, Studies for Short-Path Propagation Data and Models for Terrestrial Radiocommunication Systems in the Frequency Range 6 GHz to 100 GHz, September 2017.
[2] Recommendation ITU-R P.1238-9, Propagation Data and Prediction Methods for the Planning of Indoor Radiocommunication Systems and Radio Local Area Networks in the Frequency Range 300 MHz to 100 GHz, June 2017.
[3] Report ITU-R M.2376-0, Technical Feasibility of IMT in Bands Above 6 GHz, July 2015.
[4] Report ITU-R M.2412-0, Guidelines for Evaluation of Radio Interface Technologies for IMT-2020, October 2017.
[5] Report ITU-R M.2411-0, Requirements, Evaluation Criteria and Submission Templates for the Development of IMT-2020, November 2017.

3GPP Specifications[24]

[6] 3GPP TR 21.915, Summary of Rel-15 Work Items (Release 15), March 2019.
[7] 3GPP TR 36.873, Study on 3D Channel Model for LTE (Release 12), December 2017.
[8] 3GPP TR 37.910, Study on Self Evaluation towards IMT-2020 Submission (Release 15), December 2018.
[9] 3GPP TS 38.104, NR, Base Station (BS) Radio Transmission and Reception (Release 15), December 2018.
[10] 3GPP TS 38.201, NR, Physical Layer – General Description (Release 15), December 2018.
[11] 3GPP TS 38.202, NR, Services Provided by the Physical Layer (Release 15), December 2018.

[23] ITU-R specifications can be accessed at the following URL: http://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2020/.
[24] 3GPP specifications can be accessed at the following URL: http://www.3gpp.org/ftp/Specs/archive/.

[12] 3GPP TS 38.211, NR, Physical Channels and Modulation (Release 15), December 2018.
[13] 3GPP TS 38.212, NR, Multiplexing and Channel Coding (Release 15), December 2018.
[14] 3GPP TS 38.213, NR, Physical Layer Procedures for Control (Release 15), December 2018.
[15] 3GPP TS 38.214, NR, Physical Layer Procedures for Data (Release 15), December 2018.
[16] 3GPP TS 38.215, NR, Physical Layer Measurements (Release 15), March 2018.
[17] 3GPP TS 36.104, E-UTRA, Base Station (BS) Radio Transmission and Reception (Release 15), June 2018.
[18] 3GPP TS 38.133, NR, Requirements for Support of Radio Resource Management (Release 15), December 2018.
[19] 3GPP TS 38.300, NR, NR and NG-RAN Overall Description, Stage 2 (Release 15), December 2018.
[20] 3GPP TS 38.321, NR, Medium Access Control (MAC) Protocol Specification, (Release 15), December 2018.
[21] 3GPP TS 38.331, NR, Radio Resource Control (RRC), Protocol Specification (Release 15), December 2018.
[22] 3GPP TR 38.802, Study on New Radio Access Technology Physical Layer Aspects (Release 14), March 2017.
[23] 3GPP TR 38.812, Study on Non-Orthogonal Multiple Access (NOMA) for NR; (Release 15) October 2018.
[24] 3GPP TR 38.900, Study on Channel Model for Frequency Spectrum Above 6 GHz (Release 15), June 2018.
[25] 3GPP TR 38.901, Study on Channel Model for Frequencies From 0.5 to 100 GHz (Release 15), June 2018.

Articles, Books, White Papers, and Application Notes

[26] P. Marsch, Ö. Bulakci, 5G System Design: Architectural and Functional Considerations and Long-Term Research, Wiley, 2018.
[27] A. Zaidi, F. Athle, 5G Physical Layer: Principles, Models and Technology Components, Academic Press, 2018.
[28] E. Dahlman, S. Parkvall, 5G NR: The Next Generation Wireless Access Technology, Academic Press, 2018.
[29] F.-L. Luo, C.J. Zhang, Signal Processing for 5G: Algorithms and Implementations, Wiley-IEEE Press, 2016.
[30] S. Ahmadi, *LTE-Advanced: A Practical Systems Approach to Understanding 3GPP LTE Releases 10 and 11 Radio Access Technologies*, 2013, Academic Press.
[31] T.L. Marzetta, et al., Fundamentals of Massive MIMO, Cambridge University Press, 2016.
[32] T.S. Rappaport, R.W. Heath Jr., Millimeter Wave Wireless Communications, Prentice Hall, 2014.
[33] B. Sklar, Digital Communications: Fundamentals and Applications, Prentice Hall, 1987.
[34] F. Ademaj, et al., 3GPP 3D MIMO channel model: a holistic implementation guideline for open source simulation tools, EURASIP J. Wireless Commun. Netw. (2016) 55.
[35] B. Mondal, et al., 3D channel model in 3GPP, IEEE Commun. Mag. 53 (Issue 3) (2015).
[36] T.S. Rappaport et al., Overview of millimeter wave communications for fifth-generation (5G) wireless networks—with a focus on propagation models, IEEE Trans. Antennas Propag. (Special Issue on 5G) (2017).
[37] M. Zhang, et al., ns-3 Implementation of the 3GPP MIMO Channel Model for Frequency Spectrum above 6 GHz, ePrint of Cornell University Library, February 2017.
[38] X. Wang et al., Multicarrier Waveforms for 5G, Institut für Nachrichtenübertragung, WebDemos. <http://www.inue.uni-stuttgart.de/lehre/demo.html>.
[39] 5G New Radio, ShareTechNote. Available from: <http://www.sharetechnote.com>.
[40] D. Bharadia et al., Full duplex radios, in: Proceedings of ACM SIGCOMM, 2013.

[41] A. Sabharwal, et al., In-band full-duplex wireless: challenges and opportunities, IEEE J. Sel. Areas Commun. 32 (9) (2014).
[42] J. Campos, Understanding the 5G NR Physical Layer, Keysight Technologies, 2017.
[43] I. Chih-Lin, Seven fundamental rethinking for next-generation wireless communications, SIP 6 (2017).
[44] J. Fan, et al., Faster-than-Nyquist signaling: an overview, IEEE Access 5 (2017).
[45] A. Roessler, 5G Waveform Candidates, Application Note, Rohde & Schwarz, June 2016.
[46] B. Farhang-Boroujeny, OFDM vs. filter bank multicarrier, IEEE Signal Process Mag. (2011).
[47] White Paper, 5G Waveform & Multiple Access Techniques, Qualcomm Technologies, 2015.
[48] GTI 5G Device RF Component Research Report, 2018. Available from: <http://www.gtigroup.org>.
[49] 3GPP TSG RAN WG1, R1-165425, f-OFDM Scheme and Filter Design, 2016.
[50] Z. Ding, et al., A survey on non-orthogonal multiple access for 5G networks: research challenges and future trends, IEEE J. Sel. Areas Commun. 35 (10) (2017).
[51] Z. Wu et al., Comprehensive study and comparison on 5G NOMA schemes, IEEE Access, (2018).
[52] Y. Yuan, et al., Non-orthogonal transmission technology in LTE evolution, IEEE Commun. Mag. (2016).
[53] MediaTek, A New Era for Enhanced Mobile Broadband, White Paper, March 2018.