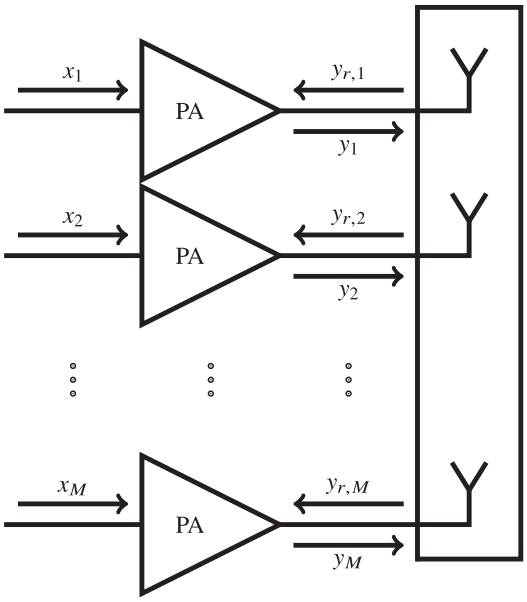


FIGURE 4.6

System model for the dual-input model. The system model for an active antenna system with nonlinear amplifiers. The reflected signal caused by mutual coupling causes a time-varying behavior which is not captured by conventional models.

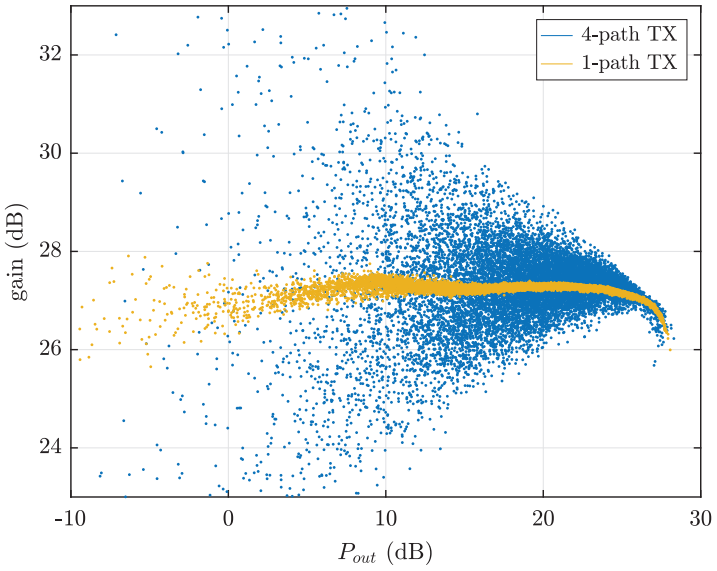


FIGURE 4.7

Measured instantaneous gain. Normalized instantaneous gain of a power amplifier with (blue; dark gray in print version) and without (yellow; light gray in print version) the impact of mutual coupling.

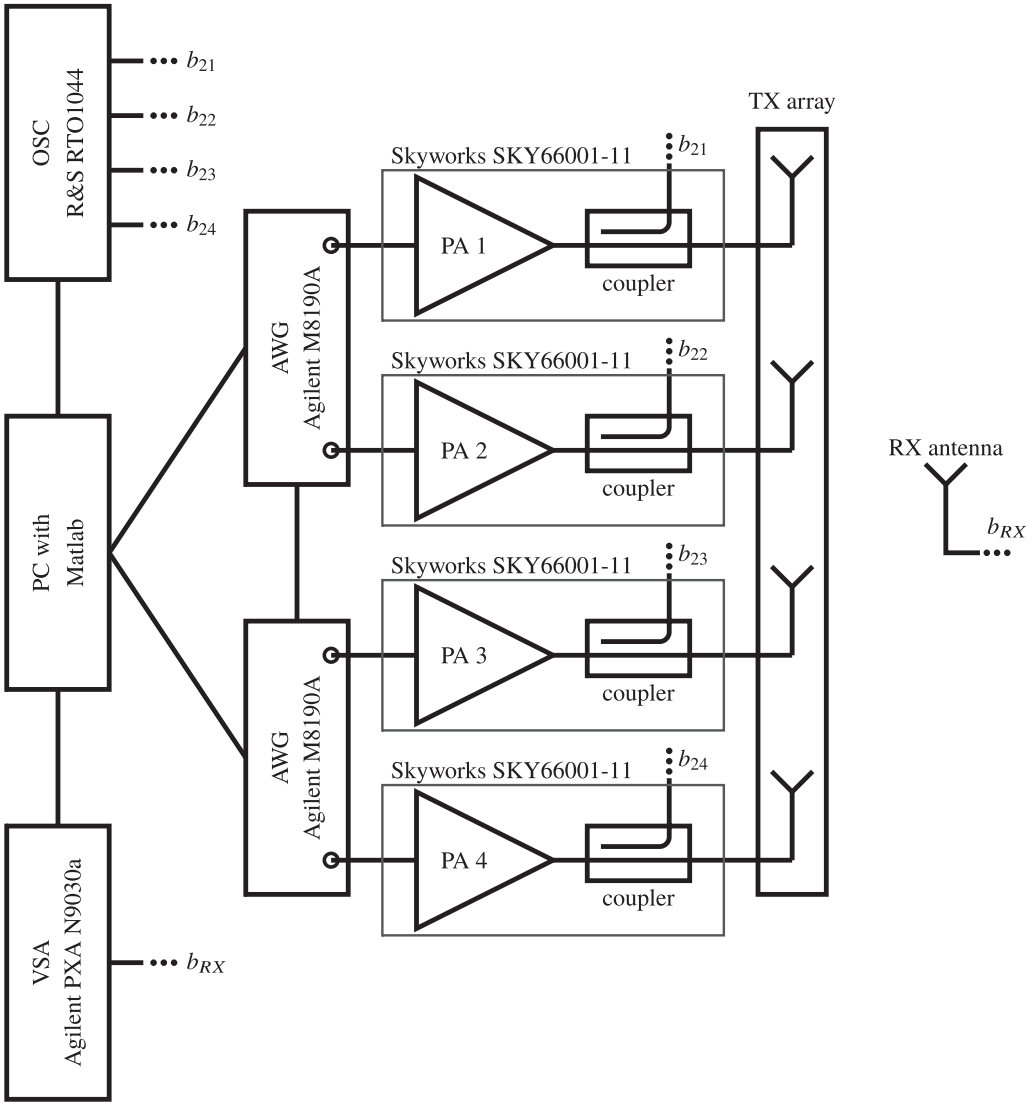


FIGURE 4.8

Measurement setup for model validation. The system model for an active antenna system with nonlinear amplifiers. The reflected signal caused by mutual coupling causes a time-varying behavior which is not captured by conventional models.

The model described in Eq. (4.24) is validated via measurements presented in [9]. These measurements were performed both conducted at the amplifier output reference plane, as well as over the air using a reference receiver. Using wide-band arbitrary waveform generators as shown in Fig. 4.8, four power amplifier modules were connected to two different configurations of a 2×2 patch antenna. The measured characteristics of these arrays are shown in Fig. 4.9. As shown, the two arrays have, via variations in patch separation, different amounts of coupling. The spectrum plots shown in Fig. 4.9 are for the case of strong mutual coupling.

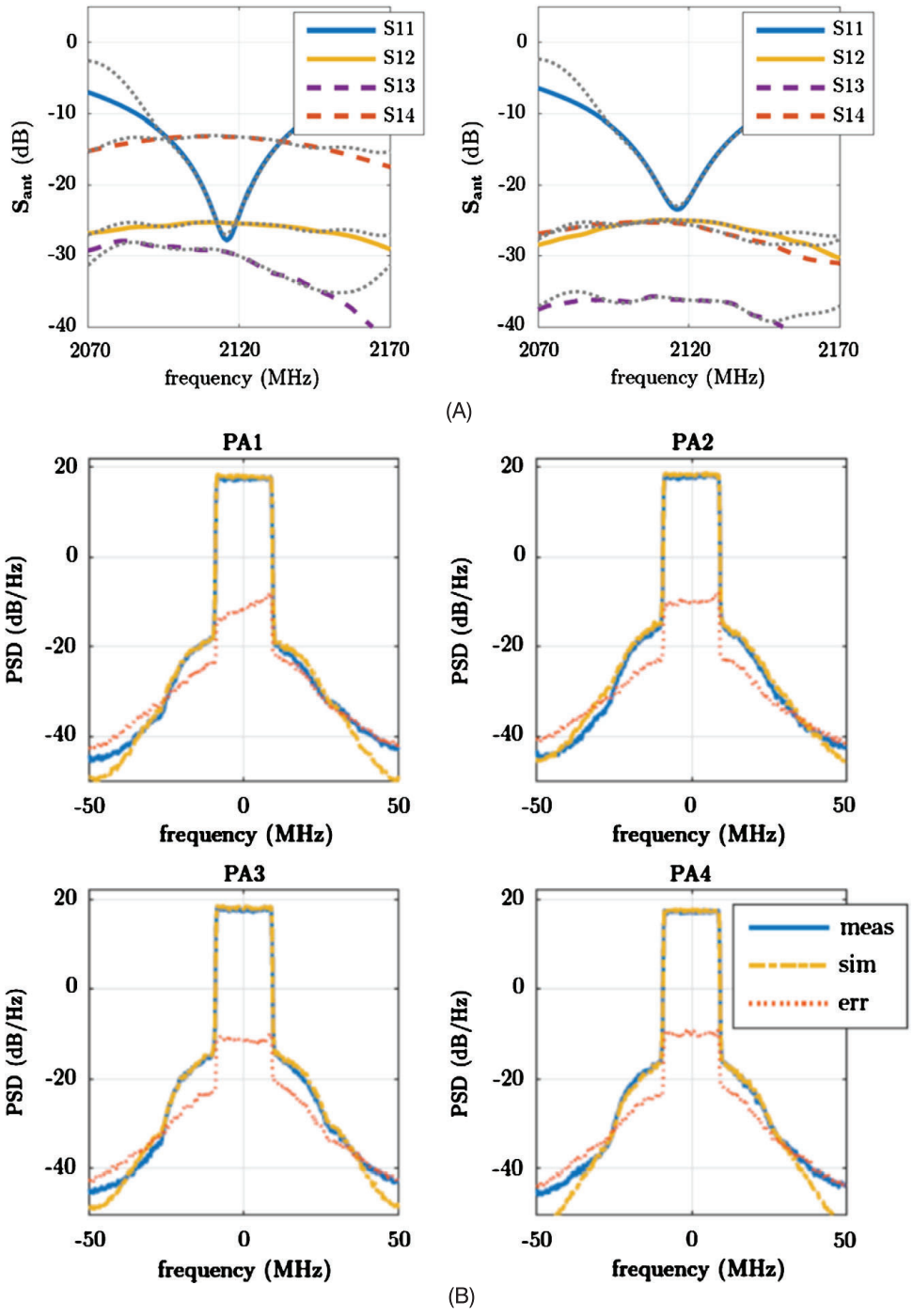


FIGURE 4.9

Antenna array characteristics and measured power spectral density (A) Measured S-parameter characteristics of the two different 2x2 patch-antenna arrays used in the validation process. To the left, the antenna spacing is 0.5λ and to the right 0.7λ, (B) measured and simulated output power spectrum from each of the four power amplifiers, followed by the model error spectrum.

4.2 OSCILLATOR PHASE NOISE

Local oscillators are the main source of phase noise in radio communication systems. It mainly stems from imperfections in both the passive and active parts of the oscillator circuitry, which in turn inflicts several types of noise sources. Some of these imperfections are:

- Thermal noise.
- Flicker or $1/f$ noise.
- Shot noise.
- Finite Q-value of resonator tank.

Together, this forms what we refer to as phase noise.

The impact of most of these sources may be mitigated by different means, for example using a PLL or simply increasing the current pumped into the resonator tank. There are drawbacks to each of these methods such as increased power consumption or increment of other noise in other parts of the spectrum. And with the trend of phase noise rapidly increasing with frequency, it is of utter importance that these effects are accurately modeled and studied. In particular, analysis of phase noise and its impact on OFDM has been performed to a very large extent.

4.2.1 PHASE-NOISE POWER SPECTRUM AND LEESON'S EQUATION

One of the early studies on phase noise conducted by Leeson in 1966, resulted in an empirical expression for phase-noise power spectral density which is commonly used to this day [16]. The power spectrum approximation is often described by

$$L(f_m) = 10\log_{10}\left[\frac{1}{2}\left(\left(\frac{f_0}{2Q_l f_m}\right)^2 + 1\right)\left(\frac{f_c}{f_m} + 1\right)\left(\frac{FkT}{P_s}\right)\right], \tag{4.25}$$

in which some relevant circuit factors are used as parameters. Here, f_0 is the operating frequency of the oscillator, Q is the loaded Q-value of the oscillator, f_m is the offset frequency at which we observe the phase noise, f_c is the $1/f$ cutoff frequency, F is the noise figure of the buffer amplifier, k is Boltzmann's constant, T is the temperature in Kelvin, and P_s is the output power of the oscillator.

Examining the power spectral density as illustrated in Fig. 4.10, the overall phase noise increases with roughly 6 dB per doubling of the center frequency. Although Leeson's equation is an empirical model, it has stood the test of time and is still commonly cited and used to predict phase-noise performance or to trade phase noise for power consumption. We will now discuss an approach used for simplistic modeling of phase noise in the presence of a phase-noise tracker.

4.2.2 PHASE-NOISE MODELING: FREE-RUNNING OSCILLATOR

As it is common to describe a communication system in discrete time due to the nature of its digital implementation, we will stay with phase-noise modeling in the discrete-time domain here. The scientific literature is however not limited to discrete time modeling as a Wiener-process may be formed in continuous time as well.

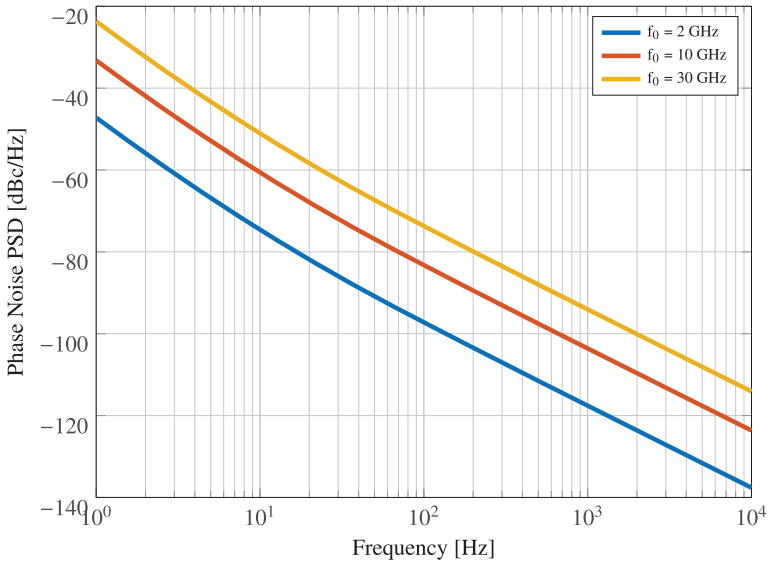


FIGURE 4.10

Phase-noise power spectral density. Phase-noise power spectral density according to Leeson's approximation for different center frequencies, $f_0 = 2$, 10, and 30 GHz.

Phase noise is multiplicative in its nature and applied as

$$y[n] = x[n]e^{i\varphi[n]}, \tag{4.26}$$

in which the phase-noise component φ is commonly modeled using a random walk process,

$$\varphi[n] = \varphi[n-1] + w[n]. \tag{4.27}$$

Just as in the continuous time case, this is an accumulative process in which the variance of φ grows over time. In the presence of a phase-noise tracker, either in the shape of a phase-locked loop or a tracker at the receiver, we may model the residual phase noise as an autoregressive process,

$$\varphi[n] = \lambda\varphi[n-1] + w[n]. \tag{4.28}$$

Here, $w \sim \mathcal{N}(0, \sigma_w^2)$ is the innovation noise and $0 < \lambda < 1$ is a constant determined mainly by the PLL. This phase-noise model is a wide-sense stationary (WSS) process which has the autocorrelation function

$$\mathbb{E}\{\varphi[n]\varphi[n-l]\} = \begin{cases} \dfrac{4\pi^2 f_c^2 t_s \kappa \lambda^{|l|}}{1-\lambda^2} & \text{if } m = m', \\ 0 & \text{else,} \end{cases} \tag{4.29}$$

where f_c is the center frequency, t_s is the sample-time, and κ is a constant depending on the quality of the oscillator and $\varphi \sim \mathcal{N}(0, 4\pi^2 f_c^2 t_s \kappa)$. Examples of the simulated power spectral density for three

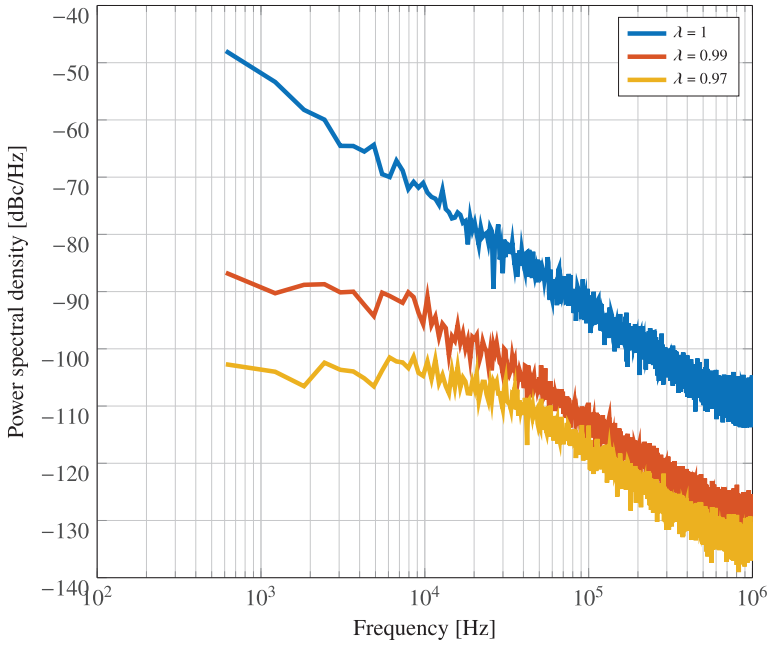


FIGURE 4.11

Phase-noise power spectral density. Simulated phase-noise spectral density using the model in (4.26)–(4.28) with three different values of λ.

different values of λ are shown in Fig. 4.11. Notice the flattening of the spectrum near the carrier frequency as λ decreases.

In practical applications, decreasing λ is analog to increasing the PLL-bandwidth which improves the phase noise around the carrier. However, this tends to increase the white noise part at higher offset frequencies, something which is not captured by this particular model.

4.2.3 PHASE-NOISE MODELING: PHASE-LOCKED LOOP

The phase noise of the PLL-based oscillator consists of three main noise sources, i.e., noises from the reference oscillator θ_{ref}, the phase–frequency detector and the loop filter θ_{LP}, and the voltage controlled oscillator (VCO) θ_{VCO}, as shown in Fig. 4.12. The Laplace transform of the phase noise of the PLL-based oscillator is given as [26]

$$\theta_{\text{out}} = \frac{N_{\text{D}} K_{\text{VCO}} Z(s) \left(K_{\text{D}} \theta_{\text{ref}} + \theta_{\text{LP}}\right) + s N_{\text{D}} \theta_{\text{VCO}}}{s N_{\text{D}} + K_{\text{D}} K_{\text{VCO}} Z(s)}, \tag{4.30}$$

where K_{D} denotes the gain of the phase–frequency detector, K_{VCO} represents the sensitivity of the VCO, $Z(s)$ represents the loop filter, and $1/N_{\text{D}}$ is the frequency divider. The noise sources include both white noise (thermal noise) and colored noise (flicker noise). The PSD of the noise source can be

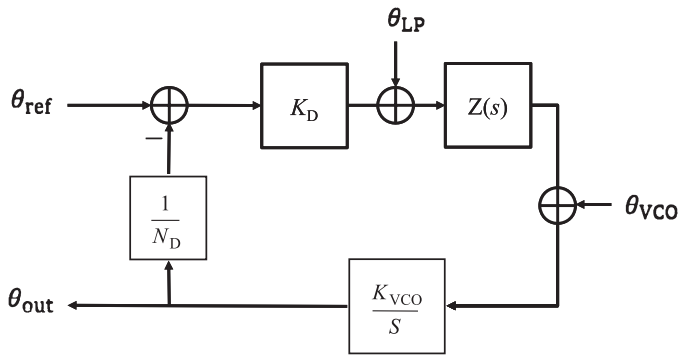


FIGURE 4.12

Phase-noise model of PLL-based oscillator. Blockdiagram of the phase-noise model for a PLL-based oscillator.

modeled as [26]

$$S(f) = S_0 \frac{1+\left(\frac{f}{f_z}\right)^2}{1+\left(\frac{f}{f_p}\right)^2} + S_0 \frac{1+\left(\frac{f_{zo}}{f_z}\right)^2}{1+\left(\frac{f_{zo}}{f_p}\right)^2} \left(1+\left(\frac{f_{zo}}{f_{po}}\right)^k\right) \frac{1}{1+\left(\frac{f}{f_{po}}\right)^k}, \tag{4.31}$$

where S_0 is the power density at zero frequency, f_p and f_z are the (3-dB) pole and corner frequencies of the thermal noise, respectively, $k = 1$ and 3 for flick noises of the PLL and of the oscillator, respectively, and f_{po} and f_{zo} are the pole and corner frequencies of the flicker noise, respectively. The detailed modeling parameters are listed in Table 4-2 of [26]. As an example, Fig. 4.13 shows the estimated PSDs of the carrier phase noise at 6, 28 and 60 GHz, respectively.

4.3 DATA CONVERTERS

Data converters are known as bottlenecks in digital communication systems as the transceiver bandwidths increases, while at the same time demands on dynamic range increase via denser modulation schemes. Amongst the sources of impairment related to data converters, the one most commonly modeled is the quantization noise. As data converters are mixed-signal components, issues with nonlinearities caused by either un-even quantization levels or output buffer amplifiers have also been modeled in the past, quite often using a Volterra series approach [4].

4.3.1 MODELING OF QUANTIZATION NOISE

The modeling of quantization noise is an area well mapped out and the study on quantization is therefore a quite mature field. Some of the most important foundations can be found in [25,23]. Part of this has led to a commonly used simplification of modeling a quantizer by adding uniform noise:

$$\mathcal{Q}(x) = x + w \tag{4.32}$$

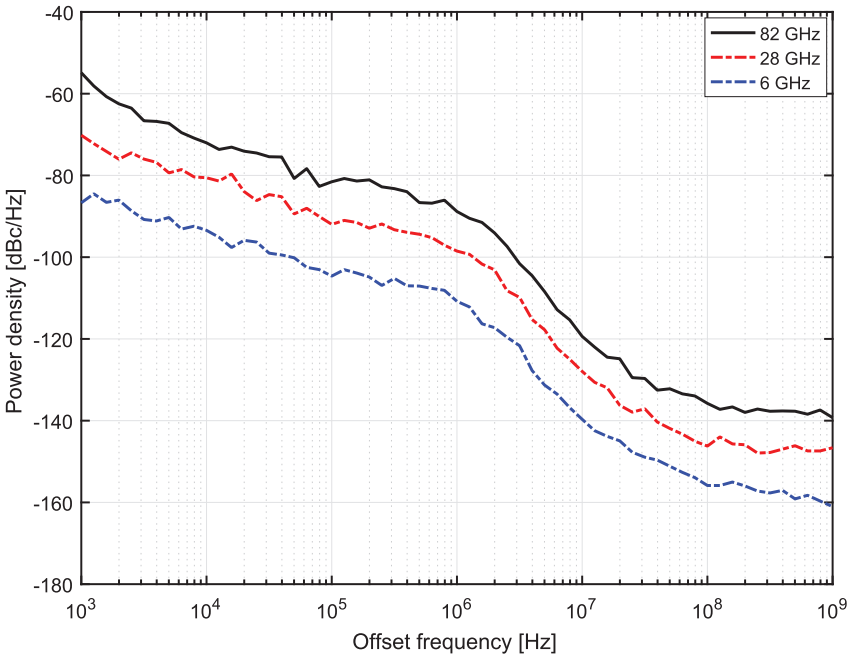


FIGURE 4.13

The PSD of the carrier phase noise. Simulated PSD of the phase noise generated by the model in Fig. 4.12, parameterized for three different center frequencies.

where $\mathcal{Q}$ denotes the quantizer function and w is an additive noise process with parameters determined by the quantizer properties and the input signal statistics. Under certain constraints, the signal to quantization noise ratio (SQNR) is modeled as

$$\text{SQNR} \approx 6.02q \tag{4.33}$$

where q is the number of bits used. This is, however, not always a good approximation depending on the input signal statistics. More exact conditions for when the quantization noise is uniform have been presented in [23]. For the general case using a DAC or ADC with q bits, we have a discrete set of 2^q possible outputs from the quantizer. Most commonly, uniform mid-rise quantizers are assumed; however, other alternatives exist. Given a finite range of the possible input signal, the quantizer can be modeled as

$$\mathcal{Q}(x) = \begin{cases} \frac{\Delta}{2}(1-2^q) & \text{if } x < -\frac{\Delta}{2}2^q, \\ \Delta\lfloor\frac{x}{\Delta}\rfloor + \frac{\Delta}{2} & \text{if } |x| < \frac{\Delta}{2}2^q, \\ \frac{\Delta}{2}(2^q-1) & \text{if } x > \frac{\Delta}{2}2^q, \end{cases} \tag{4.34}$$

in which Δ denotes the LSB size. For a quadrature signal, represented by a complex random variable, we define the input–output relation for the set of baseband DAC's or ADC's by

$$f_{\text{DAC}}(x) = \mathcal{Q}\left(\text{Re}x + i\,\text{Im}x\right) = \mathcal{Q}\left(\text{Re}x\right) + i\,\mathcal{Q}\left(\text{Im}x\right). \tag{4.35}$$

As exact modeling of the quantization noise may be difficult and linear, stochastic approaches as proposed in [10] may be sufficient, but we will leave the discussion for Section 4.7.

4.4 STATISTICAL MODELING

In certain applications such as system-level performance assessment, it may be difficult to make use of behavioral models as they often operate on oversampled time-domain signals. In most communication system engineering applications, the signals are modeled as stochastic random variables with given second-order statistics described with a covariance matrix. As such a signal passes through a nonlinear function, for example a building block, the statistical properties changes. With advanced multiantenna systems making use of the spatial properties of the communication channel, the importance of knowing the second-order statistics increases as it determines the spatial properties of both the transmit signal as well as the hardware impairments.

4.4.1 THE BUSSGANG THEOREM AND THE SYSTEM MODEL

Determining the second-order statistics of the distortion is straightforward using the Bussgang theorem. Let $f(\cdot)$ be a nonlinear function and $x \sim \mathcal{N}(0, \sigma_x^2)$. The input–output cross-covariance is equal to the input autocorrelation up to a constant C [5], as

$$R_{f(x),x} = C R_x. \tag{4.36}$$

This property may be used in order to formulate a linear stochastic model which describes the output of a nonlinear system when the input is modeled using a Gaussian random variable x. The model is written as

$$f(x) = \alpha x + d \tag{4.37}$$

where $d \sim \mathcal{N}(0, \sigma_d^2)$ is the distortion noise which is orthogonal to the scaled input signal αx and α is the gain constant, determined by

$$\alpha = \frac{\mathbb{E}\left[x \bar{y}\right]}{\sigma_x^2}. \tag{4.38}$$

From this, we have $\mathbb{E}\left[\alpha x \bar{d}\right] = 0$. For the multiantenna case, considering a system with M antennas, we may model this in matrix-vector format as

$$\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{e} \tag{4.39}$$

where $\mathbf{G}$ is the complex gain matrix and $\mathbf{e}$ is the distortion noise. The gain matrix is defined as

$$\mathbf{G} = \operatorname{diag}\left(\mathbb{E}\left[x_1 \bar{y}_1\right], \ldots, \mathbb{E}\left[x_M \bar{y}_M\right]\right). \tag{4.40}$$

The remaining distortion noise, which is orthogonal to $\mathbf{Gx}$, is then distributed as

$$\mathbf{e} \sim \mathbb{C}\mathcal{N}\left(\mathbf{0}, \mathbf{C}_\mathbf{e}\right) \tag{4.41}$$

where

$$\mathbf{C_e} = \mathbf{C_y} - \mathbf{G}\mathbf{C_x}\mathbf{G}^H, \tag{4.42}$$

which forces $\mathbf{e}$ and $\mathbf{x}_n$ to be mutually orthogonal. This description is equivalent to the model discussed in Section 4.1.5 as they have exactly the same properties, which stems from using mutually orthogonal basis functions in the shape of Itô–Hermite polynomials.

4.5 STOCHASTIC MODELING OF POWER AMPLIFIERS

In the case of power amplifiers, derivation of the distortion covariance is straightforward by repeatedly using the moment theorem for Gaussians [21]. This may, however, be tedious, as the number of cross-products and higher-order moments tends to explode. For this reason, we will stick to the canonical third-order model here. Using the common third-order polynomial as foundation, written as

$$y[n] = \theta_1 x[n] + \theta_2 x[n]|x[n]|^2, \tag{4.43}$$

we would, as illustrated in [10], be able to derive the model in the following form:

$$\mathbf{y} \quad = \quad \mathbf{G}^{\text{PA}}\mathbf{x} + \mathbf{e}^{\text{PA}}. \tag{4.44}$$

Here $\mathbf{G}^{\text{PA}} = \mathbf{I}_{MN} \otimes \text{diag}\left(\mathbf{g}^{\text{PA}}\right) \in \mathbb{C}^{MN \times MN}$, or more explicitly

$$\mathbf{G}^{\text{PA}} = \mathbf{I}_{MN} \otimes \left(\theta_1 \mathbf{I}_{MN} + 2\theta_2 \text{diag}(\mathbf{C_x})\right). \tag{4.45}$$

The distortion covariance due to orthogonality will become

$$\mathbf{C}_{\mathbf{e}^{\text{PA}}} = \mathbf{C_y} - \mathbf{G}^{\text{PA}}\mathbf{C_x}(\mathbf{G}^{\text{PA}})^H \tag{4.46}$$

in which $\mathbf{G}^{\text{PA}}$ is given by Eq. (4.45), $\mathbf{C_x}$ is the transmit covariance matrix which is given by precoder design and $\mathbf{C_y}$ is the covariance matrix of the distorted output. This has to be manually computed using the moment theorem for Gaussians, which in the end results in

$$\begin{aligned} \mathbf{C_y} \quad &= \quad |\theta_1|\mathbf{C_x} + 2|\theta_2|^2\mathbf{C_x} \circ |\mathbf{C_x}|^2 \\ &+ \quad 2\bar{\theta}_1\theta_2 \left(\mathbf{I}_{MN} \otimes \text{diag}(\mathbf{C}_{\mathbf{x}_n})\right)\mathbf{C_x} + 2\bar{\theta}_2\theta_1\mathbf{C_x}\left(\mathbf{I}_{MN} \otimes \text{diag}(\mathbf{C}_{\mathbf{x}_n})\right) \\ &+ \quad 2|\theta_2|^2 \left(\mathbf{I}_{MN} \otimes \text{diag}(\mathbf{C}_{\mathbf{x}_n})\right) \times \mathbf{C_x}\left(\mathbf{I}_{MN} \otimes \text{diag}(\mathbf{C}_{\mathbf{x}_n})\right) \end{aligned} \tag{4.47}$$

where $\circ$ is the element-wise product. A full derivation of this can be found in [10].

4.6 OSCILLATOR PHASE NOISE

In the case of phase noise, formulating a stochastic model in the same format as for the power amplifier is not trivial as phase noise is in itself not generated through a nonlinear function but is in fact already

described as a stochastic process. Therefore, we may not use the same straightforward method using the Bussgang theorem. However, we may do a signal decomposition using LMMSE Estimation in order to formulate a model in the linear same format.

We recall that the behavioral model for the phase noise is described by

$$y[n] = x[n]e^{i\varphi[n]} \tag{4.48}$$

where $\varphi[n] = \lambda\varphi[n-1] + w[n]$. The stochastic model for the phase noise, in this case, is written as

$$\mathbf{x} = \mathbf{G}^{\text{PN}}\mathbf{w} + \mathbf{e}^{\text{PN}} \tag{4.49}$$

where

$$\mathbf{G}^{\text{PN}} = e^{-\frac{2\pi^2 f_c^2 t_s \kappa}{1-\lambda^2}}\mathbf{I}_{MN}. \tag{4.50}$$

In the case of phase noise with separate oscillators using PLL synchronization, the covariance of the error-term becomes

$$\mathbf{C}_{\mathbf{e}^{\text{PN}}} = \mathbf{C}_{\mathbf{y}} - e^{-\frac{4\pi^2 f_c^2 t_s \kappa}{1-\lambda^2}}\mathbf{C}_{\mathbf{x}} \tag{4.51}$$

in which the output covariance is

$$\mathbf{C}_{\mathbf{y}} = \text{diag}\left(\boldsymbol{\psi}\right)\text{diag}\left(\mathbf{C}_{\mathbf{x}}\right) + e^{-\frac{2\pi^2 f_c^2 t_s \kappa}{1-\lambda^2}}\text{nondiag}\left(\mathbf{C}_{\mathbf{x}}\right). \tag{4.52}$$

Here, $\boldsymbol{\psi} = \text{vec}\left(\left[\boldsymbol{\psi}_0, \ldots, \boldsymbol{\psi}_{N-1}\right]\right)$ which for each $\boldsymbol{\psi}_m$ may be computed as

$$\boldsymbol{\psi}_m = e^{-\frac{4\pi^2 f_c^2 t_s \kappa}{1-\lambda^2}\left(1-\lambda^{\left|(m+\frac{N}{2})\text{mod}N-\frac{N}{2}\right|}\right)}. \tag{4.53}$$

It is of interest to notice that the gain and covariance matrix depend only on PLL, the oscillator parameter λ, and κ.

4.7 STOCHASTIC MODELING OF DATA CONVERTERS

Finally, before moving on to concatenate the model-blocks into a complete transmitter model, we need a description of the quantization noise from the data converters. Just as with power amplifier distortion and phase noise, we model the process in a linear fashion. We have

$$\mathbf{x} \quad = \quad \mathbf{G}^{\text{DAC}}\mathbf{w} + \mathbf{e}^{\text{DAC}}. \tag{4.54}$$

The gain matrix for a DAC, denoted

$$\mathbf{G}^{\text{DAC}} = \mathbf{I}_{MN} \otimes \text{diag}\left(\mathbf{g}^{\text{DAC}}\right), \tag{4.55}$$

will element-wise be computed as [11]

$$\mathbf{g}^{\text{DAC}} = \frac{\Delta}{\sqrt{\pi}}\text{diag}\left(\mathbf{C}_{\mathbf{x}_n}\right)^{-\frac{1}{2}} \sum_{i=1}^{2^q-1} e^{-\Delta^2(i-2^{q-1})^2\left(\mathbf{C}_{\mathbf{x}_n}\right)^{-1}}. \tag{4.56}$$

Now again, following the orthogonality we get the error covariance matrix:

$$\mathbf{C}_{\mathbf{e}^{\text{DAC}}} = \mathbf{C}_{\mathbf{y}} - \mathbf{G}^{\text{DAC}}\mathbf{C}_{\mathbf{x}}\mathbf{G}^{\text{DAC}}. \tag{4.57}$$

The quantized output covariance matrix, from [11, Eq. 10], is then computed as

$$\begin{aligned}\mathbf{C}_{\mathbf{y}} \approx\ & \frac{\Delta^2}{2}(2^q-1)^2\mathbf{I}_{BN} - 4\Delta^2 \sum_{i=1}^{2^q-1}\left(i-2^{q-1}\right) \\ \times\ & \Phi\left(\sqrt{2}\text{diag}\left(\mathbf{C}_{\mathbf{x}}\right)^{-\frac{1}{2}}\left(i-2^{q-1}\right)\right) \\ +\ & \mathbf{G}^{\text{DAC}}\text{nondiag}\left(\mathbf{C}_{\mathbf{x}}\right)\mathbf{G}^{\text{DAC}}.\end{aligned} \tag{4.58}$$

With the individual models for each of the subsystem component, a more complete transmitter model may be assembled.

4.8 MODEL CONCATENATION AND SIMULATIONS

To assemble the chain forming a complete transmitter chain model, we may use the linear properties of each subsystem model to our advantage. If we consider the linear models for two different (vector) nonlinear systems

$$\begin{aligned}\mathbf{y}_1 &= \mathbf{G}_1\mathbf{x} + \mathbf{e}_1 \\ \mathbf{y}_2 &= \mathbf{G}_2\mathbf{x} + \mathbf{e}_2\end{aligned} \tag{4.59}$$

the concatenated model for the series non-linearity becomes

$$\mathbf{y} = \mathbf{G}_2(\mathbf{G}_1\mathbf{x} + \mathbf{e}_1) + \mathbf{e}_2. \tag{4.60}$$

Using this step repeatedly, we may find the transmitter-model as derived in [10]

$$\begin{aligned}\mathbf{y} &= \mathbf{G}^{\text{PA}}\left(\mathbf{G}^{\text{PN}}\left(\mathbf{G}^{\text{DAC}}\mathbf{z} + \mathbf{e}^{\text{DAC}}\right) + \mathbf{e}^{\text{PN}}\right)\mathbf{w} + \mathbf{e}^{\text{PA}} \\ &= \underbrace{\mathbf{G}^{\text{PA}}\mathbf{G}^{\text{PN}}\mathbf{G}^{\text{DAC}}\mathbf{z}}_{\text{Scaled and rotated signal}} + \underbrace{\mathbf{G}^{\text{PA}}\mathbf{G}^{\text{PN}}\mathbf{e}^{\text{DAC}} + \mathbf{G}^{\text{PA}}\mathbf{e}^{\text{PN}} + \mathbf{e}^{\text{PA}}}_{\text{Distortion noise}},\end{aligned} \tag{4.61}$$

from which we may observe an equivalent gain matrix, $\mathbf{G}^{\text{PA}}\mathbf{G}^{\text{PN}}\mathbf{G}^{\text{DAC}}$ and an accumulative noise process with rescaled contributions from each subsystem component. Using the concatenated model, we may now describe the signal-to-interference, distortion and noise ratio (SINDR).

4.8.1 SIGNAL-TO-INTERFERENCE AND NOISE RATIO

As described in more detail in [10], the SINDR for an OFDM system at the kth UE and the lth subcarrier may be approximated by means of the concatenated model discussed in previous section. The SINDR is

$$\gamma_{k,l} \approx \frac{\exp\left(-\frac{2\pi f_c t_s \kappa}{1-\lambda^2}\right)\left|\hat{\mathbf{h}}_{k,l}^T \operatorname{diag}(\mathbf{g}^{\mathrm{PA}}) \operatorname{diag}(\mathbf{g}^{\mathrm{DAC}})\hat{\mathbf{p}}_{k,l}^T\right|^2}{I_{k,l}+E_{k,l}+N_0} \tag{4.62}$$

in which we have that

$$I_{k,l} = \exp\left(-\frac{2\pi f_c t_s \kappa}{1-\lambda^2}\right)\sum_{v\neq k}\left|\hat{\mathbf{h}}_{k,l}^T \operatorname{diag}(\mathbf{g}^{\mathrm{PA}}) \operatorname{diag}(\mathbf{g}^{\mathrm{DAC}})\hat{\mathbf{p}}_{v,l}^T\right|^2 \tag{4.63}$$

is the interference power and

$$\begin{aligned} E_{k,l} = \Bigg[&(\mathbf{F}_N \otimes \mathbf{I}_M)\left(\mathbf{G}^{\mathrm{PA}}\left(\exp\left(-\frac{2\pi f_c t_s \kappa}{1-\lambda^2}\right)\times\left(\mathbf{C}_{\mathbf{v}}-\mathbf{G}^{\mathrm{DAC}}\mathbf{C}_{\mathbf{z}}\mathbf{G}^{\mathrm{DAC}}\right)\right.\right. \\ &\left.\left.+\left(\mathbf{C}_{\mathbf{w}}-\exp\left(-\frac{2\pi f_c t_s \kappa}{1-\lambda^2}\right)\mathbf{C}_{\mathbf{v}}\right)\left(\mathbf{G}^{\mathrm{DAC}}\right)^H+\mathbf{C}_{\mathbf{x}}-\mathbf{G}^{\mathrm{DAC}}\mathbf{C}_{\mathbf{w}}\left(\mathbf{G}^{\mathrm{DAC}}\right)^H\right)\left(\mathbf{I}_M\otimes\mathbf{F}_N^H\right)\Bigg]_{k+lK,k+lK} \end{aligned} \tag{4.64}$$

is the received distortion power on the lth subcarrier and the kth user. Here, $\mathbf{F}_N$ is the $N \times N$ DFT-matrix for which $\mathbf{F}_N\mathbf{F}_N^H = \mathbf{I}_N$.

Looking at the interference power, comparing to conventional signal-to-interference-and-noise ratio (SINR)-analysis of massive MU-MIMO systems, further degeneration stemming from non-ideal hardware may be observed.

4.8.2 SIMULATIONS

In order to illustrate the usage of the stochastic transceiver model discussed in this chapter, we will present some basic link-simulations. These make use of the stochastic models in order to emulate an OFDM-based massive MU-MIMO array transmitter and we compare the results with simulations using conventional models. We will consider an OFDM-system with frequency selective linear precoding using maximum ratio transmission (MRT) and zero-forcing (ZF). The precoding matrices are defined as

$$\mathbf{P}_k^{\mathrm{MRT}} = \alpha_{\mathrm{MRT}}\frac{1}{M}\mathbf{H}_k^H \tag{4.65}$$

$$\mathbf{P}_k^{\mathrm{ZF}} = \alpha_{\mathrm{ZF}}\frac{1}{M}\mathbf{H}_k\left(\mathbf{H}_k\mathbf{H}_k^H\right)^{-1} \tag{4.66}$$

where

$$\mathbf{H}_k = \sum_{t=0}^{T}\hat{\mathbf{H}}_t e^{-jk\frac{2\pi}{N}t} \tag{4.67}$$

is the $M \times K$ frequency-domain matrix describing the channel at subcarrier k. α_{MRT} and α_{ZF} are chosen such that $\mathbb{E}\left[|\mathbf{x}_n|^2\right] = 1$ for the precoded vector $\mathbf{x}_n$. The full set of parameters are found in Table 4.2. We consider an OFDM-based MU-MIMO system with M antennas and K users and a Rayleigh inde-

Table 4.2 Simulation parameters

Parameter	**Setting**
Carrier frequency	3 GHz
Carrier type	OFDM
Number of occupied subcarriers	1200
FFT size	4096
Carrier bandwidth	20 MHz
Channel parameter	**Setting**
Channel type	Rayleigh IID in spherical coordinates
Number of channel taps	5
Power delay profile	Exponential
DAC parameter	**Setting**
q	6 bits
LSB (Δ)	0.0081
Phase-noise parameter	**Setting**
λ	0.99
κ	$5 \cdot 10^{-17}$
Power amplifier parameter	**Setting**
θ_1	1
θ_2	$-0.03491 + \mathrm{i}0.00565$

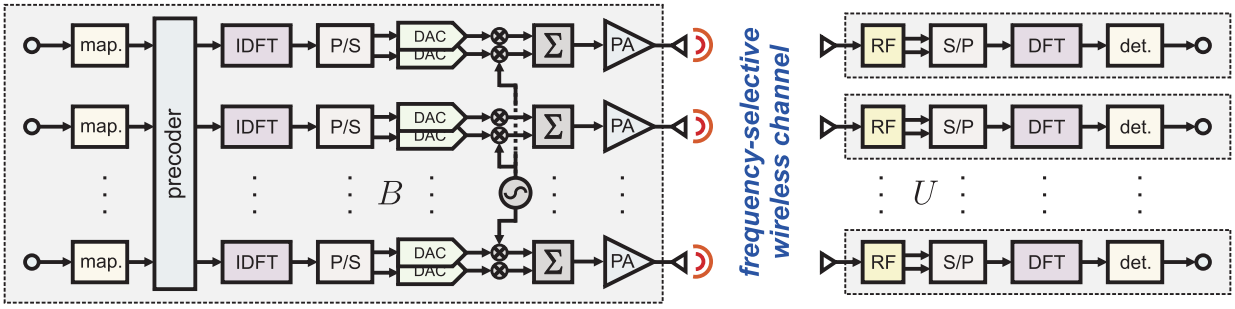


FIGURE 4.14

Complex baseband system model. The complex baseband system model including a base-station equipped with M antennas, each consisting of DAC's, local oscillators and power amplifiers. This is followed by the radio channel and K served users.

pendent identically distributed (IID) channel. Fig. 4.14 shows the system model including impairment models for quantization noise, phase noise and power amplifier distortion.

4.8.3 SIMULATION RESULTS

The simulation results are summarized in Figs. 4.15–4.17. Fig. 4.15 shows the simulated power spectral density of each model block as well as the composite transmitter model, compared to the corresponding

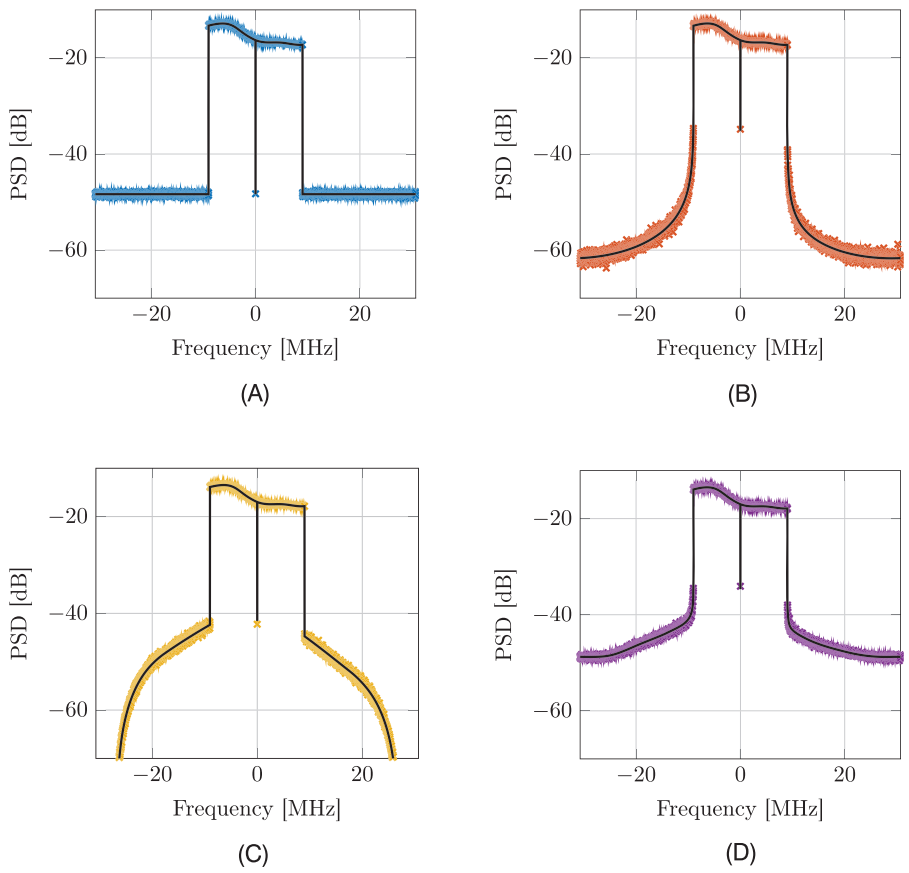


FIGURE 4.15

Output power spectral density from the statistical models. PSD of the transmitted signal for the case when each hardware impairment in the RF-chain are considered separately, and when all of the hardware impairments are considered together. The markers correspond to simulation results whereas the black lines are the analytical outcome. (A) Finite-resolution DACs only (ideal LO and linear PA). (B) Nonideal LO only (infinite-resolution DAC and linear PA). (C) Nonlinear PAs only (infinite-resolution DACs and ideal LO). (D) Concatenated impairment model (finite-resolution DACs, nonideal LO, and nonlinear PA).

behavioral models. As may be seen, the stochastic model accurately predicts the power spectral density in not only each model block, but also in the case of the complete transmit model. The PSD is accurately predicted both in- and out-of band, which is otherwise not the case for most linear and stochastic models which today is widely used in the literature, [3].

Fig. 4.16 shows the radiated power of both the transmit signal (A) and the adjacent channel power (B). Here we may spot similar effects to those noted in Section 4.1.5. In a similar manner, the inband distortion takes on a similar radiation pattern as the transmit signal in this case.

Finally, Fig. 4.17 shows the performance in terms of uncoded BER with and without impairment models. As may be observed, the radio hardware impairments has a negative impact on the uncoded BER performance for both ZF and MRT.

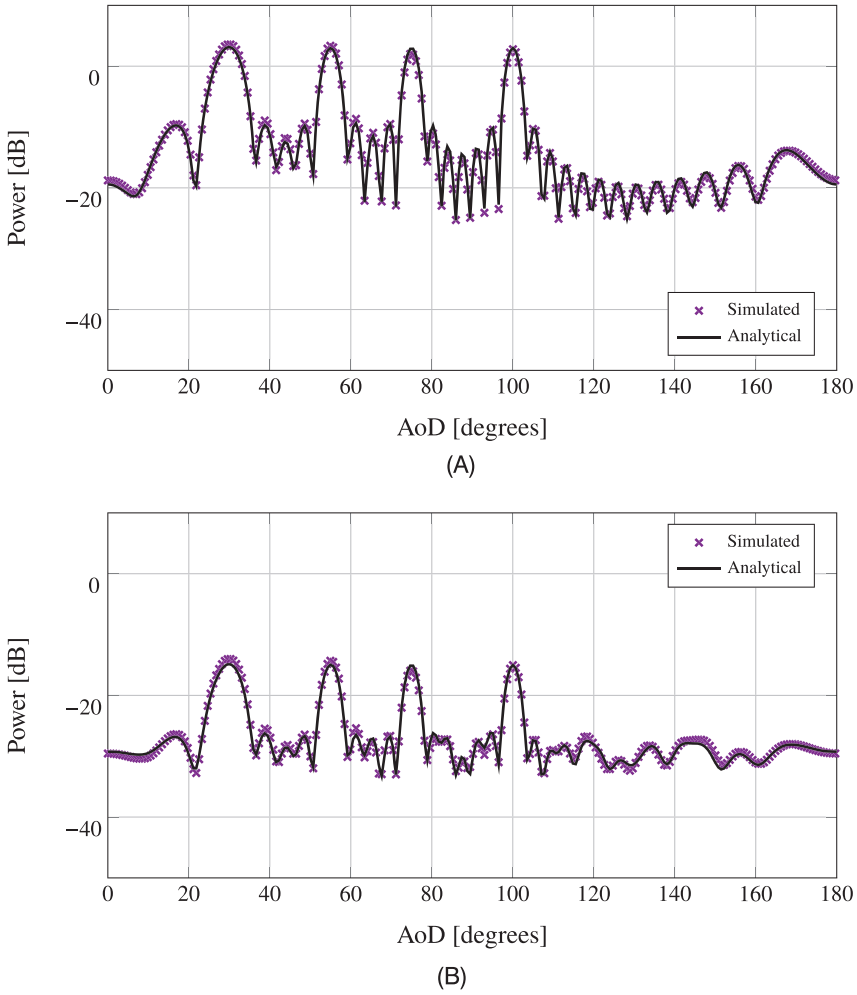


FIGURE 4.16

Simulated radiation patterns using stochastic impairment models. Simulated radiation patterns of the desired signal (A) and the impairment component (B), consisting of power amplifier distortion, phase noise, and quantization noise.

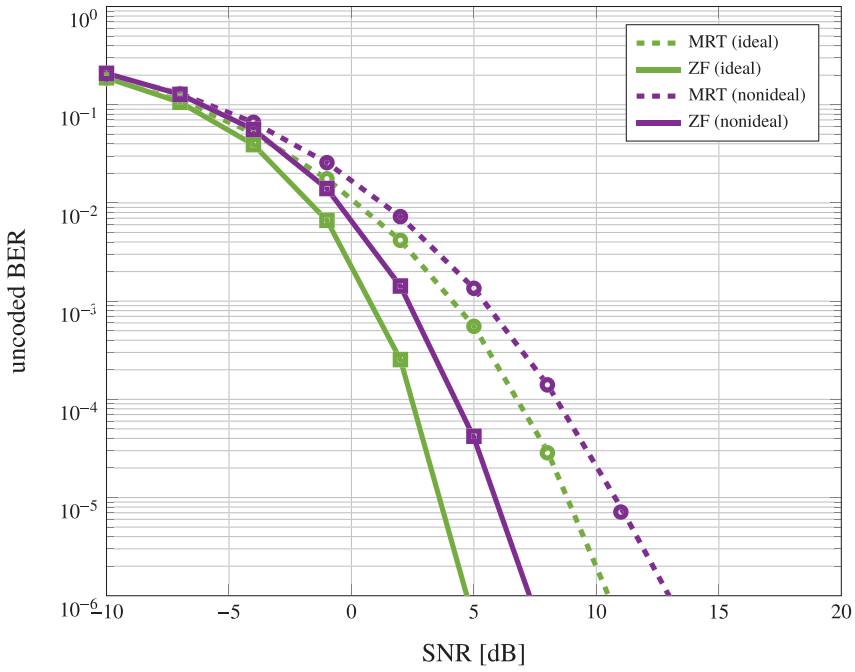


FIGURE 4.17

Link-performance assessment using stochastic impairment models. Comparison of uncoded BER using linear precoding, with and without the impact of hardware impairments.

REFERENCES

[1] S. Afsardoost, T. Eriksson, C. Fager, Digital predistortion using a vector-switched model, IEEE Transactions on Microwave Theory and Techniques (ISSN 0018-9480) 60 (4) (2012, April) 1166–1174, https://doi.org/10.1109/TMTT.2012.2184295.

[2] S. Benedetto, E. Biglieri, R. Daffara, Modeling and performance evaluation of nonlinear satellite links-a Volterra series approach, IEEE Transactions on Aerospace and Electronic Systems (ISSN 0018-9251) AES-15 (4) (1979, July) 494–507, https://doi.org/10.1109/TAES.1979.308734.

[3] E. Björnson, J. Hoydis, M. Kountouris, M. Debbah, Massive MIMO systems with non-ideal hardware: energy efficiency, estimation, and capacity limits, IEEE Transactions on Information Theory 60 (11) (2014) 7112–7139.

[4] N. Bjorsell, P. Suchánek, P. Handel, D. Ronnow, Measuring Volterra kernels of analog-to-digital converters using a stepped three-tone scan, IEEE Transactions on Instrumentation and Measurement 57 (4) (2008) 666–671.

[5] J.J. Bussgang, Crosscorrelation functions of amplitude-distorted gaussian signals, 1952.

[6] H. Chireix, High power outphasing modulation, Proceedings of the Institute of Radio Engineers (ISSN 0731-5996) 23 (11) (1935, Nov.) 1370–1392, https://doi.org/10.1109/JRPROC.1935.227299.

[7] L. Ding, G.T. Zhou, Effects of even-order nonlinear terms on predistortion linearization, in: Digital Signal Processing Workshop, 2002 and the 2nd Signal Processing Education Workshop. Proceedings of 2002 IEEE 10th, 2002, pp. 1–6.

[8] W.H. Doherty, A new high efficiency power amplifier for modulated waves, Proceedings of the Institute of Radio Engineers (ISSN 0731-5996) 24 (9) (1936, Sept.) 1163–1182, https://doi.org/10.1109/JRPROC.1936.228468.

[9] K. Hausmair, S. Gustafsson, C. Sánchez-Pérez, P.N. Landin, U. Gustavsson, T. Eriksson, C. Fager, Prediction of nonlinear distortion in wideband active antenna arrays, IEEE Transactions on Microwave Theory and Techniques (ISSN 1557-9670) 65 (11) (2017, Nov.) 4550–4563, https://doi.org/10.1109/TMTT.2017.2699962.

[10] S. Jacobsson, U. Gustavsson, D. Astely, G. Durisi, C. Studer, T. Eriksson, Modeling and analysis of transmit-RF impairments in massive MU-MIMO OFDM downlink, 2018, in preparation.

[11] S. Jacobsson, G. Durisi, M. Coldrey, C. Studer, Linear precoding with low-resolution DACs for massive MU-MIMO-OFDM downlink, CoRR, arXiv:1709.04846, http://arxiv.org/abs/1709.04846, 2017.

[12] J. Kim, K. Konstantinou, Digital predistortion of wideband signals based on power amplifier model with memory, Electronics Letters (ISSN 0013-5194) 37 (23) (2001, Nov.) 1417–1418, https://doi.org/10.1049/el:20010940.

[13] P.N. Landin, D. Rönnow, RF PA modeling considering odd–even and odd order polynomials, in: Communications and Vehicular Technology in the Benelux (SCVT), 2015 IEEE Symposium on, 2015, pp. 1–6.

[14] P.N. Landin, S. Gustafsson, C. Fager, T. Eriksson, Weblab: a web-based setup for PA digital predistortion and characterization [application notes], IEEE Microwave Magazine (ISSN 1527-3342) 16 (1) (2015, Feb.) 138–140, https://doi.org/10.1109/MMM.2014.2367857.

[15] P.M. Lavrador, T.R. Cunha, P.M. Cabral, J.C. Pedro, The linearity-efficiency compromise, IEEE Microwave Magazine (ISSN 1527-3342) 11 (5) (2010, Aug.) 44–58, https://doi.org/10.1109/MMM.2010.937100.

[16] D.B. Leeson, A simple model of feedback oscillator noise spectrum, Proceedings of the IEEE (ISSN 0018-9219) 54 (2) (1966, Feb.) 329–330, https://doi.org/10.1109/PROC.1966.4682.

[17] C. Mollén, U. Gustavsson, T. Eriksson, E.G. Larsson, Spatial characteristics of distortion radiated from antenna arrays with transceiver nonlinearities, IEEE Transactions on Wireless Communications (2018), accepted for publication, arXiv:1711.02439, http://arxiv.org/abs/1711.02439, 2017.

[18] C. Mollén, U. Gustavsson, T. Eriksson, E.G. Larsson, Out-of-band radiation measure for MIMO arrays with beamformed transmission, in: 2016 IEEE International Conference on Communications (ICC), 2016, May, pp. 1–6.

[19] G.E. Moore, Cramming more components onto integrated circuits, Electronics 38 (8) (1965) 114.

[20] D.R. Morgan, Z. Ma, J. Kim, M.G. Zierdt, J. Pastalan, A generalized memory polynomial model for digital predistortion of RF power amplifiers, IEEE Transactions on Signal Processing 54 (10) (2006) 3852–3860.

[21] I.S. Reed, On a moment theorem for complex Gaussian processes, IEEE Transactions on Information Theory 8 (3) (1962, Apr.) 194–195, https://doi.org/10.1109/TIT.1962.1057719.

[22] D. Root, et al., Polyharmonic distortion modeling, IEEE Microwave Magazine 7 (3) (2006) 44–57.

[23] A. Sripad, D. Snyder, A necessary and sufficient condition for quantization errors to be uniform and white, IEEE Transactions on Acoustics, Speech, and Signal Processing (ISSN 0096-3518) 25 (5) (1977, Oct.) 442–448, https://doi.org/10.1109/TASSP.1977.1162977.

[24] V. Volterra, Sopra le funzioni che dipendono da altre funzioni, Tip. della R. Accademia dei Lincei, 1887.

[25] B. Widrow, I. Kollár, Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications, Cambridge University Press, Cambridge, UK, ISBN 9780521886710, 2008.

[26] P. Zetterberg, A. Wolfgang, A. Westlund, et al., Initial multi-node and antenna transmitter and receiver architectures and schemes, mmMAGIC Deliverale D5.1, 2016.

CHAPTER

MULTICARRIER WAVEFORMS 5

The waveform is a core component of any communication technology. Broadly speaking, there are two main categories of waveforms: i) single-carrier waveforms, ii) multicarrier waveforms. 2G and 3G cellular systems (GSM, UMTS, CDMA2000) as well as ZigBee and Bluetooth adopted single-carrier waveforms. The 4G cellular system (LTE) employs a multicarrier waveform. Multicarrier waveforms are also used in asymmetric digital subscriber line (ADSL), wireless local area network (WLAN), WiMAX, digital audio broadcast (DAB), and digital video broadcast-terrestrial (DVB-T) standards. Typically, single-carrier waveforms have a low peak-to-average power ratio (PAPR), which makes them power efficient—suitable for coverage limited scenarios and extending battery life of user equipment. On the other hand, multicarrier waveforms provide high spectral efficiency, flexible resource allocation in the frequency domain, and possibly easy integration with multiantenna technology. These are the key drivers for 5G NR.

NR will support various use cases with different deployments in frequencies from below 1 GHz to 100 GHz. Although single-carrier waveforms can be interesting for massive IoT devices (with extended battery life) and for operation in high carrier frequencies (where transmission loss is high), the multicarrier waveforms have been seen as main candidates of 5G due to the reasons mentioned above. The 3GPP evaluated several multicarrier waveforms as well as single carrier waveforms. It was observed that different waveforms have their pros and cons, which makes different waveforms suitable in different scenarios. However, the 3GPP also acknowledged the fact that the 5G NR radio interface based on multiple waveforms will make the overall system design complex. Considering overall performance, system requirements, and the need for a single waveform, the 3GPP concluded in favor of OFDM for both uplink and downlink transmissions. Moreover, the DFT-Spread OFDM (DFTS-OFDM) waveform, which has single-carrier properties, has been kept as an option for uplink transmission in coverage limited scenarios.

This chapter presents various multicarrier waveforms and provides a comparison that has led to the down selection of OFDM for 5G NR. The chapter is structured as follows. Section 5.1 introduces state-of-the-art OFDM-based and filter bank multicarrier (FBMC)-based waveforms and Section 5.2 introduces the DFTS-OFDM waveform. Section 5.3 describes the waveform design requirements for NR. Section 5.4 presents the key indicators for waveform comparisons, based on which Section 5.5 compares performance of the multicarrier waveforms.

5.1 MULTICARRIER WAVEFORMS

The main principle behind multicarrier waveforms is to split a high rate data stream into multiple low rate streams that are transmitted simultaneously over a number of carriers (termed subcarriers). By doing so, the symbol duration for each low rate subcarrier increases and therefore the relative amount of time dispersion caused by a multipath (frequency selective) channel decreases. This makes

5G Physical Layer. https://doi.org/10.1016/B978-0-12-814578-4.00010-2
Copyright © 2018 Elsevier Ltd. All rights reserved.

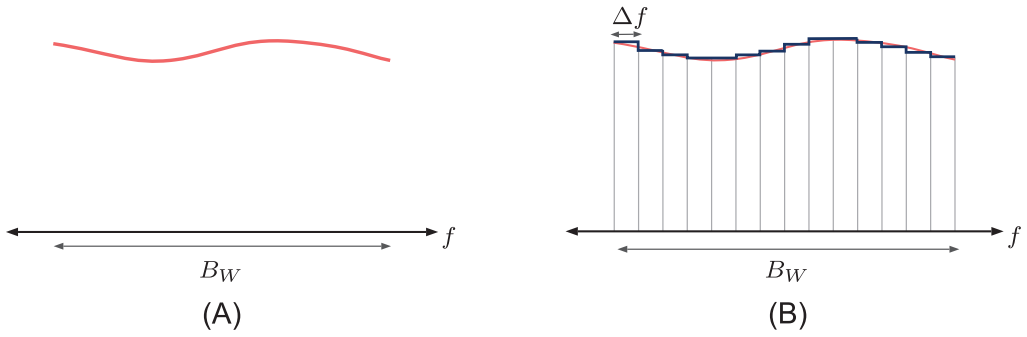


FIGURE 5.1

(A) Frequency-selective channel. (B) Frequency flat subchannels.

a multicarrier waveform robust to the intersymbol interference caused by the multipath channel. Let us look at this in the frequency domain. A frequency-selective fading channel is split into multiple (approximately) frequency flat channels and each subcarrier now experiences frequency flat fading, as illustrated in Fig. 5.1. This avoids complex time domain equalization, making the multicarrier transmissions suitable for high data rate wireless communication.

Multicarrier waveforms provide the freedom of multiplexing multiple users in the frequency domain (allocating different subcarriers to different users) and spatial domain (mapping different subcarriers to different antennas). Furthermore, multicarrier waveforms are robust to narrowband interference compared to single-carrier waveforms. In a multicarrier system, narrowband interference only corrupts a small number of subcarriers, whereas in a single-carrier transmission an entire link can fail due to narrowband interference.

5.1.1 THE PRINCIPLE OF ORTHOGONALITY

Mathematically, a multicarrier waveform in baseband can be expressed as

$$\widetilde{x}(t) = \sum_{s\in\mathbb{Z}} \sum_{i\in I} X_{s,i}\ \widetilde{p}_i(t - sT), \tag{5.1}$$

where s denotes the waveform symbol index, $\mathbb{Z}$ is a set of integers, i denotes the subcarrier index, $I \in \{i = 1, 2, \ldots, N\}$ contains indices of the subcarriers, N is the total number of subcarriers, $X_{i,s} \in \mathbb{C}$ is a complex modulated (QAM or phase shift keying (PSK)) symbol mapped to the subcarrier i and the symbol index s, and $\widetilde{p}_i(t)$ is a pulse shape (or prototype filter) used for the subcarrier i. Typically, $\widetilde{p}_i(t)$ is a bandpass filter centered at the subcarrier frequency f_i. For example, for OFDM, $\widetilde{p}_i(t) = \mathrm{e}^{j2\pi f_i t}$ for $0 \le t \le T$ and $\widetilde{p}_i(t) = 0$ elsewhere. In a multicarrier system, the number of active subcarriers is usually chosen smaller than the total number of subcarriers to relax filtering operations, i.e., $X_{i,s} = 0$ for some i at the band edges. The modulation and demodulation operations for a multicarrier system are illustrated in Fig. 5.2.

The demodulation operation is based on matched filtering, which is optimal for linear and additive noise channels in the sense of maximizing signal-to-noise ratio (SNR) of the demodulated (reconstructed) signal. The matched filter demodulation comprises the convolution of the received signal $\widetilde{r}(t)$

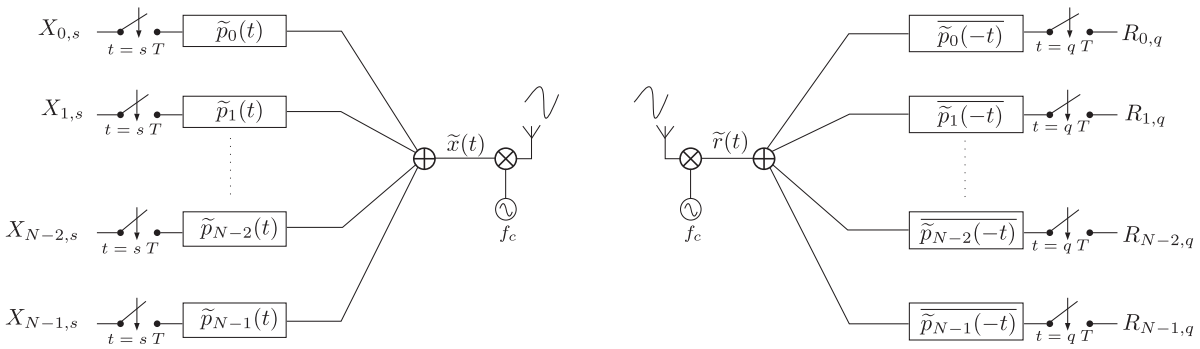


FIGURE 5.2

Modulation and demodulation process in a multicarrier system.

with the complex conjugated time-reversed version of the transmitter prototype filter (or pulse shape), followed by sampling at a rate equal to $\frac{1}{T}$. For simplicity, let us assume that $\widetilde{r}(t) = \widetilde{x}(t)$ and denote $R_{l,q}$ as demodulated (reconstructed) QAM (or PSK) symbol from a waveform symbol q and a subcarrier l. The demodulation operation is then given by

$$\begin{aligned} R_{l,q} &= \widetilde{r}(t) * \overline{\widetilde{p}_l(-t)}\Big|_{t=qT} \\ &\overset{(a)}{=} \left\langle \widetilde{r}(t), \widetilde{p}_l(t - qT) \right\rangle \\ &\overset{(b)}{=} \sum_{s\in\mathbb{Z}} \sum_{i\in I} X_{s,i} \left\langle \widetilde{p}_i(t - sT), \widetilde{p}_l(t - qT) \right\rangle, \end{aligned} \tag{5.2}$$

where $\overline{\widetilde{p}_i(t)}$ denotes complex conjugation of $\widetilde{p}_i(t)$; $*$ is the convolution operator; $\langle v_1(t), v_2(t)\rangle := \int_{\infty}^{\infty} v_1(t) v_2{}^*(t) dt$ denotes the inner product between $v_1(t)$ and $v_2(t)$; (a) follows from the definition of the convolution operation (i.e., expressing convolution in terms of correlation); and (b) follows from the assumption that $\widetilde{r}(t) = \widetilde{x}(t)$. From (5.2), we can see that the perfect reconstruction of the transmitted QAM symbols ($R_{l,q} = X_{l,q}$) requires the following orthogonality condition to be fulfilled:

$$\left\langle \widetilde{p}_i(t - sT), \widetilde{p}_l(t - qT) \right\rangle = \delta_{i,l}\, \delta_{s,q}, \tag{5.3}$$

where $\delta_{a,b}$ is the Kronecker delta function, defined as $\delta_{a,b} = 1$ when $a = b$ and $\delta_{a,b} = 0$ when $a \neq b$.

A multicarrier waveform that satisfies the orthogonality condition in (5.3) is called an orthogonal waveform; for example, the orthogonal frequency division multiplexing (OFDM) waveform. Let us look at the prototype filters used for OFDM and see how they fulfill the orthogonality condition. For an ith subcarrier in OFDM, the prototype filter is given by

$$\widetilde{p}_i(t) := \widetilde{p}(t)\, \mathrm{e}^{j2\pi f_i t}\,, \tag{5.4}$$

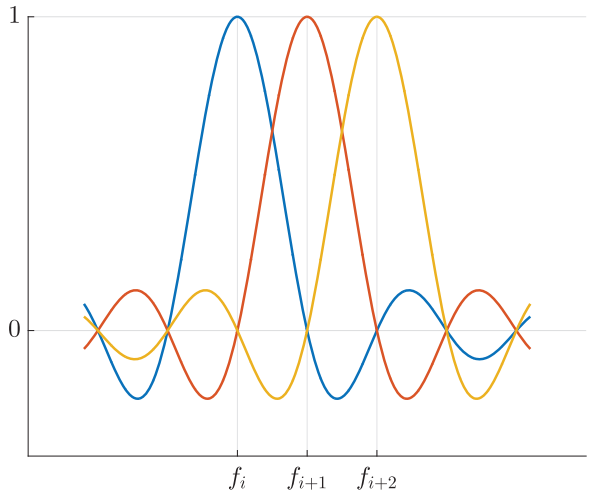


FIGURE 5.3

Frequency domain illustration of OFDM subcarriers.

where $f_i = \frac{i}{T}$ and $\widetilde{p}(t)$ is a rectangular filter with length T defined as

$$\widetilde{p}(t) := \begin{cases} \frac{1}{\sqrt{T}} & \text{if } 0 \le t \le T \\ 0 & \text{elsewhere} \end{cases}, \tag{5.5}$$

where the scaling $\frac{1}{\sqrt{T}}$ is simply chosen to get unit energy, i.e., $\int_{-\infty}^{+\infty} |\widetilde{p}(t)|^2 \, dt = 1$. Without loss of generality, the inner product (cf. (5.3)) between the prototype filters is given by

$$\begin{aligned} \langle \widetilde{p}_i(t), \widetilde{p}_l(t) \rangle &= \int_0^T \widetilde{p}_i(t) \, \overline{\widetilde{p}_l(t)} \, dt \\ &= \int_0^T \widetilde{p}(t) \, \mathrm{e}^{j2\pi f_i t} \, \widetilde{p}(t) \, \mathrm{e}^{-j2\pi f_l t} \, dt \\ &= \int_0^T \mathrm{e}^{j2\pi (f_i - f_l) t} \, dt \\ &= \int_0^T \mathrm{e}^{j2\pi \frac{i-l}{T} t} \, dt, \end{aligned}$$

which is nonzero (and equal to 1) if and only if $i = l$. We observe that it is the selection of $f_i = \frac{i}{T}$ that makes the OFDM waveform orthogonal. A frequency domain illustration of the OFDM subcarriers is shown in Fig. 5.3. The spacing between consecutive subcarriers is $f_{i+1} - f_i = \frac{i+1}{T} - \frac{i}{T} = \frac{1}{T} := \Delta f$, i.e., the subcarrier spacing Δf is the inverse of the symbol duration T.

When the orthogonality condition is not fulfilled, the waveform is nonorthogonal, for example, the FBMC waveforms. In the FBMC waveforms, the prototype filters of the subcarriers are not orthogonal within the same symbol as well as between the adjacent symbols. Typically, there is interference be-

tween subcarriers of the same symbols and between subcarriers of the adjacent different symbols (due to overlapping waveform symbols in the time domain). The inner product (based on (5.3)) takes the following form:

$$\left\langle \widetilde{p}_i(t-sT), \widetilde{p}_l(t-qT) \right\rangle = \begin{cases} 1 & \text{if } i=l \text{ and } s=q \\ \varepsilon_{i,l} & \text{if } i\neq l \text{ and } s=q, \\ \beta_{i,l,s,q} & \text{else,} \end{cases} \tag{5.6}$$

where $\varepsilon_{i,l}$ corresponds to an interference between subcarriers of the same symbol and $\beta_{i,l,s,q}$ is associated with an interference between subcarriers of two different symbols. These two types of interference are known as self-interferences—$\varepsilon_{i,l}$ is referred to as self-intercarrier interference and $\beta_{i,l,s,q}$ is referred to as self-intersymbol interference. The structure of the FBMC waveforms will be discussed in Section 5.1.3.

5.1.2 OFDM-BASED WAVEFORMS

There are several OFDM-based waveforms. In the following, we will present four major variants of the OFDM waveform, namely, CP-OFDM, windowed OFDM (W-OFDM), filtered OFDM (F-OFDM), and universally filtered OFDM (UF-OFDM).

A general discrete-time OFDM-based waveform is given by

$$x[n] = \left(\sum_{s\in\mathbb{Z}} \sum_{i\in I} X_{i,s}\ p_i[n-sN_{sc}] \right) * w_{tx}[n], \tag{5.7}$$

where I contains indices of the active subcarriers,[1] N_{sc} is the total number of subcarriers, $X_{i,s} \in \mathbb{C}$ is the modulated symbol corresponding to the subcarrier i and the time index s, $p_i[n] := p[n]\,\mathrm{e}^{j2\pi \frac{i}{N_{sc}} n}$ is the prototype filter (or pulse shape) used for the subcarrier i, and $w_{tx}[n]$ is a filter to improve spectral characteristics of the transmitted signal.[2] The modulation and demodulation operations of the OFDM-based waveforms are shown in Fig. 5.4 and Fig. 5.5, respectively.

5.1.2.1 Cyclic Prefix OFDM

CP-OFDM employs a rectangular pulse shape, i.e., $p[n]$ in (5.7) is given by

$$p[n] := \begin{cases} \frac{1}{\sqrt{N_{sc}}} & \text{if } 0 \le n \le N_{sc}-1, \\ 0 & \text{elsewhere,} \end{cases} \tag{5.8}$$

where N_{sc} is the total number of subcarriers. There is no post filtering operation, i.e., $w_{tx}[n] = \delta[n]$ in (5.7). To combat the intersymbol interference (ISI) in a multipath channel, a cyclic prefix (CP)

[1] In a multicarrier system, the number of active subcarriers is usually chosen smaller than the total number of subcarriers to relax filtering operations.

[2] The sharper the spectrum roll-off of the signal, the easier it is to fulfill the out-of-band (OOB) emission requirements specified by the 3GPP for cellular systems. The OOB emission requirements are specified in terms of a spectrum emission mask and an adjacent channel leakage ratio for both base stations and devices.

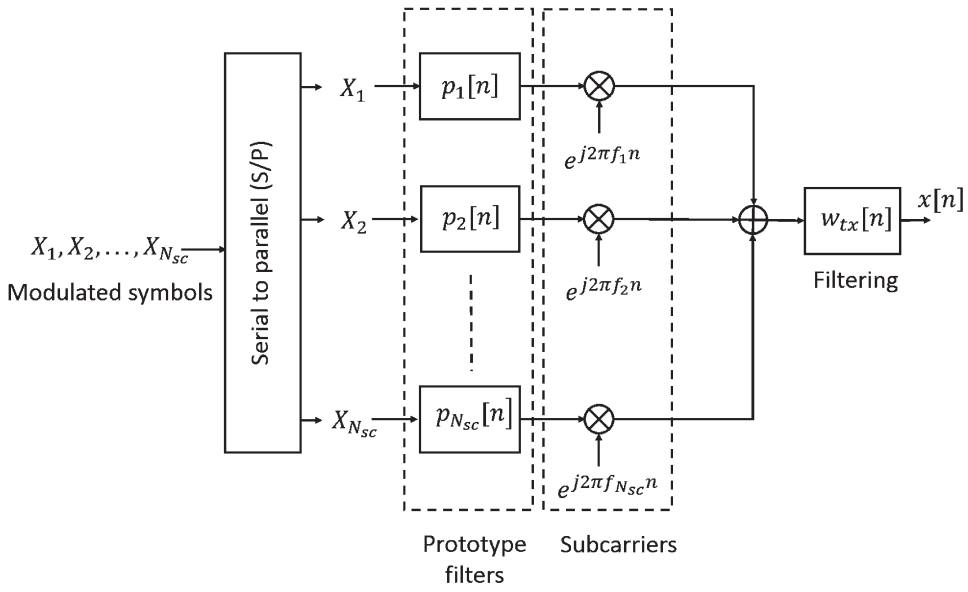


FIGURE 5.4

A general modulation structure of OFDM waveforms.

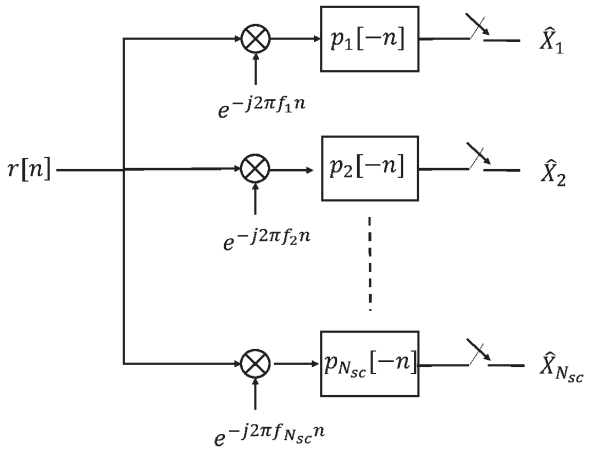


FIGURE 5.5

A general demodulation structure of OFDM waveforms.

is inserted in the OFDM symbol. The CP refers to the cyclic extension of an OFDM symbol, that is, appending the last N_{cp} samples of the OFDM symbol to the front of the symbol as illustrated in Fig. 5.6. If the CP duration is chosen greater than the delay spread of the channel, then the received OFDM signal does not suffer from any ISI. In LTE, the normal CP duration is 7% of the core OFDM symbol (i.e., $N_{sc} = 2048$ samples and $N_{cp} = 144$ samples).

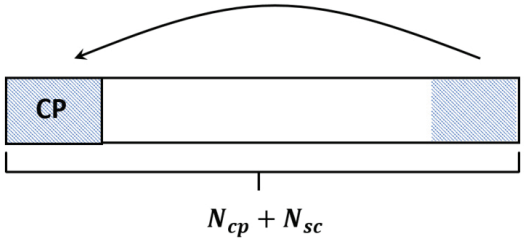


FIGURE 5.6

Cyclic prefix is inserted to avoid intersymbol interference due to the multipath channel.

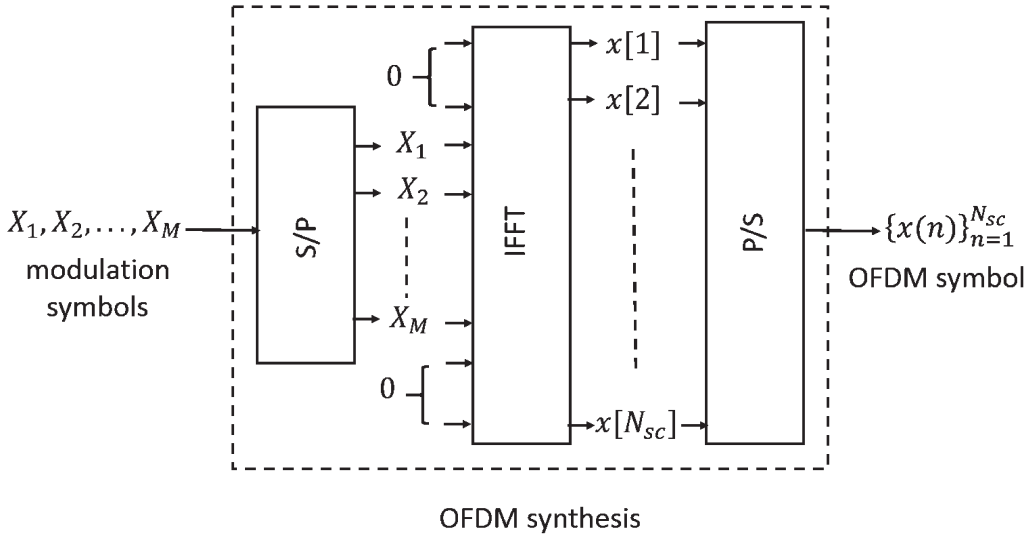


FIGURE 5.7

OFDM waveform is efficiently synthesized via fast Fourier transform.

In practice, CP-OFDM modulator is implemented via the computationally efficient fast Fourier transform (FFT). As shown in Fig. 5.7, M modulated QAM symbols ($\{X_i\}_{i=1}^{M}$) are mapped to orthogonal subcarriers via a serial-to-parallel (S/P) transformation followed by an IFFT of size N and parallel-to-serial (P/S) conversion. In this example, the total number of subcarriers is N and the number of active subcarriers is equal to M. In practice, the number of active subcarriers is kept smaller than the total number of subcarriers to relax filtering operations. This is achieved by zero padding QAM symbols prior to the IFFT operation, as shown in Fig. 5.7.

5.1.2.2 Windowed OFDM

The spectrum of CP-OFDM decays rather slowly. The main reason for its slow decay is the signal discontinuities at the symbol boundaries due to the rectangular pulse shape used in OFDM (cf. (5.8)). To improve the spectral shape of OFDM, a nonrectangular pulse shape $p[n]$ with smooth edges can be employed in (5.7), which we refer to as windowed-OFDM (W-OFDM) waveform. As an example,

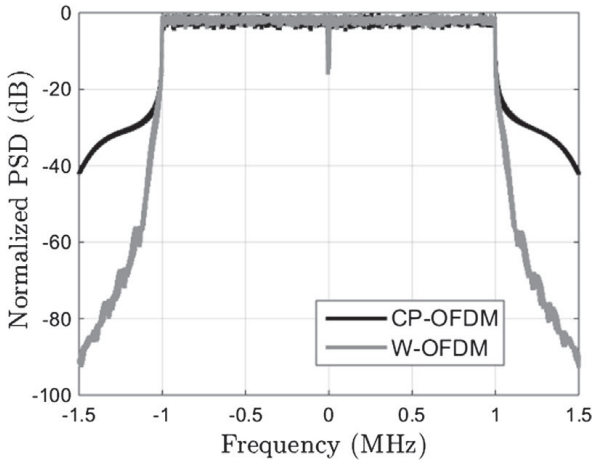


FIGURE 5.8

PSDs of CP-OFDM and W-OFDM waveforms.

the spectrum of W-OFDM (considering a window with linear slopes) is compared with CP-OFDM in Fig. 5.8. We observe that W-OFDM has a much sharper spectrum roll-off.[3]

The windowing can be done either at the transmitter side (to suppress interference to adjacent bands) or at the receiver side (to suppress interference pick-up from adjacent bands) [10,14]. In the following, we explain the transmitter side and the receiver side windowing. The transmitter side windowing is illustrated in Fig. 5.9. In the transmitter side windowing, the boundaries of each OFDM symbol are multiplied with a smooth slope in the time domain, increasing smoothly from 0 to 1 (increasing slope) or decreasing smoothly from 1 to 0 (decreasing slope). The increasing slope is applied at the beginning of the CP, while the decreasing slope is applied after the end of the core OFDM symbol within an extra added cyclic suffix. Fig. 5.9 also shows that the increasing slope of the next OFDM symbol overlaps with the decreasing slope of the previous OFDM symbol. Since the receiver only keeps the samples of the core OFDM symbol, the transmitter side windowing is transparent to the receiver. Next, we discuss the receiver side windowing.

A standard OFDM receiver cuts out the desired OFDM symbol period by applying a rectangular window in the time domain to the received signal and then subsequently applies the FFT for demodulating the subcarriers. Application of the rectangular window in the time domain corresponds to a convolution in the frequency domain with a sinc function. The sinc-like function leads to a high interference pick-up from any adjacent nonorthogonal signals such as OFDM signals with other numerologies. To reduce the interference pick-up, the rectangular window must be replaced by a smooth window function. The receiver side windowing operation is illustrated in Fig. 5.10. A smooth increasing window slope is applied at the boundary between the CP and the core OFDM symbol (half within each); a decreasing smooth window slope is applied at the boundary between the core OFDM symbol and the added cyclic suffix. If the applied window slopes fulfill the Nyquist criterion (i.e., they are

[3]The sharper the spectrum roll-off of the signal, the easier it is to fulfill the out-of-band emission requirements specified by the 3GPP for cellular systems.

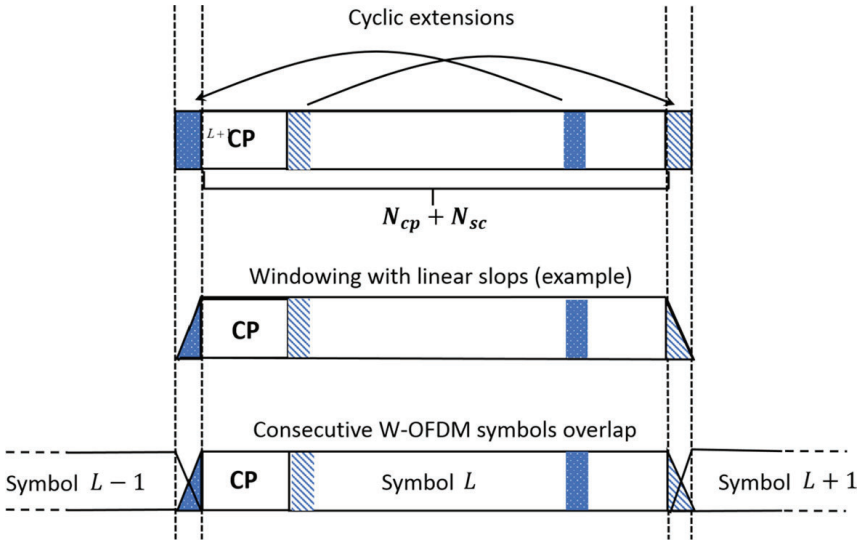


FIGURE 5.9

The transmitter side windowing.

center symmetric), the signal part cut away by the decreasing windowing slope is the same as the remaining signal part after application of the increasing window slope within the CP, since the CP is a copy of the last part of the OFDM symbol. If the windowed CP part is added to the last part of the core OFDM symbol, the core OFDM symbol is restored at its second boundary. The core OFDM symbol can also be restored at the first symbol boundary by applying the same trick. Now the complete OFDM is restored, and its subcarriers are orthogonal again. The FFT is applied to the restored core OFDM symbol as indicated in Fig. 5.10.

5.1.2.3 Filtered OFDM

Like windowing (or pulse shaping), filtering is another technique to improve spectral characteristics or reduce out-of-band (OOB) leakage of OFDM waveform [16]. Filtering (like windowing) is typically done at both transmitter and receiver sides to reduce interference to adjacent bands and reduce interference pick-up from adjacent bands, respectively. Let us refer to the general OFDM waveform modulation structure in (5.7). In filtered OFDM (F-OFDM), the prototype filter (pulse shape) $p[n]$ is rectangular including the CP, as we have in CP-OFDM. In addition, a filter $w_{tx}[n]$ is employed to suppress OOB leakage (cf. Fig. 5.4). The filter $w_{tx}[n]$ is derived as the product of an ideal bandpass filter impulse response and a time domain mask (or a window), that is,

$$w_{tx}[n] = h[n]w[n], \tag{5.9}$$

where $h[n]$ is an ideal bandpass filter with the selected signal bandwidth and $w[n]$ is a window function. Various window functions are common for the construction of FIR filters with different characteristics, for example, Hanning, Hamming, Blackman, Bartlett, and Kaiser window functions.

The filter $w_{tx}[n]$ is dependent on the signal bandwidth, which implies that the filter need to be dynamically designed or selected based on the signal bandwidth. This is different from W-OFDM,

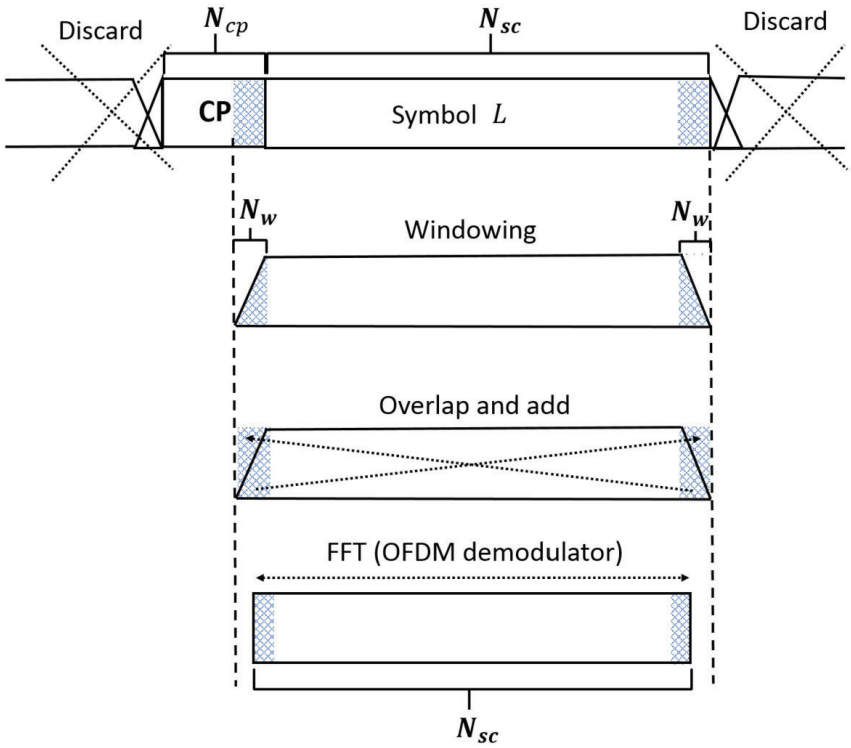


FIGURE 5.10

The receiver side windowing.

whose construction does not change with the signal bandwidth (i.e., the pulse shape $p[n]$ does not change with signal bandwidth). Another aspect of F-OFDM is that when long filters are used for an improved OOB suppression, there can be significant group delay due to the filtering. For example, if filters with length equal to half OFDM symbol are used at the transmitter and at the receiver sides, then the filter processing delay is equal to one OFDM symbol. The large processing delay increases overhead for TDD link direction switching, which is not suitable for latency critical applications.

5.1.2.4 Universally Filtered OFDM

UF-OFDM waveform is synthesized by performing filtering on groups of subcarriers (subbands) to improve the spectral shape (reduce OOB leakage) of each subcarrier group (subband) [12,9]. The subcarrier group-wise (subband) filtering is motivated by the fact that the scheduling operation in cellular systems (e.g., LTE and NR) is done based on subbands (resource blocks[4]). The subband filtering improves spectral characteristics. As an example, the spectrum of W-OFDM is compared with CP-OFDM in Fig. 5.11.

The transmitter structure of UF-OFDM is illustrated in Fig. 5.12. We assume that there are M active subcarriers and the subband filtering is performed on groups of B subcarriers (i.e., each subband

[4]In 4G LTE and 5G NR, a resource block comprises of 12 subcarriers.

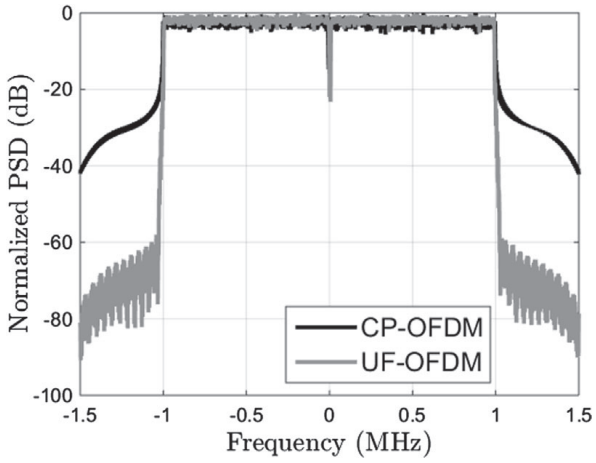


FIGURE 5.11

PSDs of CP-OFDM and UF-OFDM waveforms.

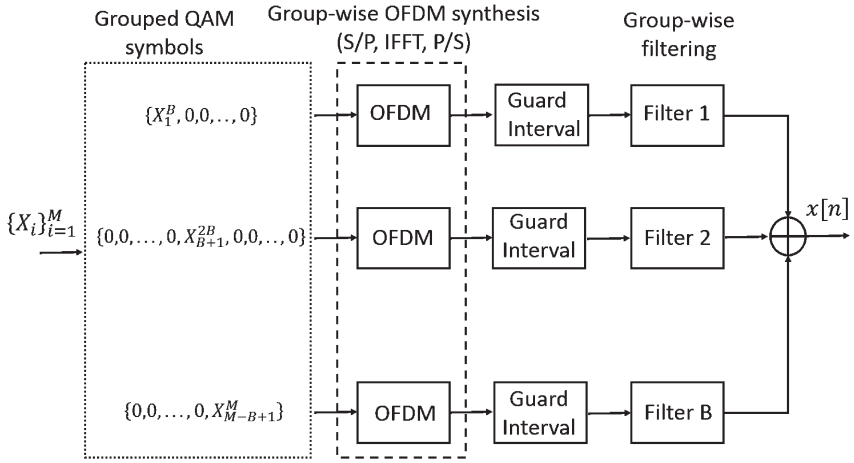


FIGURE 5.12

UF-OFDM transmitter structure.

has B subcarriers). As shown in Fig. 5.12, a set of QAM modulated symbols $\{X_1, X_2, \ldots, X_M\}$ are first divided into groups and zero padded. Each zero padded group follows the same process as we have for the OFDM synthesis, i.e., serial-to-parallel conversion, N_{sc}-point IFFT, and parallel-to-serial conversion. Afterwards, each symbol is padded with zeros in the time domain (i.e., a guard interval (GI) is inserted for each symbol) and filtered. The GI is inserted to prevent intersymbol interference due to the transmit filter delay. For an ISI free UF-OFDM synthesis, the GI length should be at-least equal to the filter length, as illustrated in Fig. 5.13.

The receiver processing of the UF-OFDM waveform is shown in Fig. 5.14. UF-OFDM waveform does not contain the CP, therefore, the cyclic convolution property is not preserved. Unlike the CP-OFDM demodulator that discards the CP part of the symbol before the FFT operation, the UF-OFDM

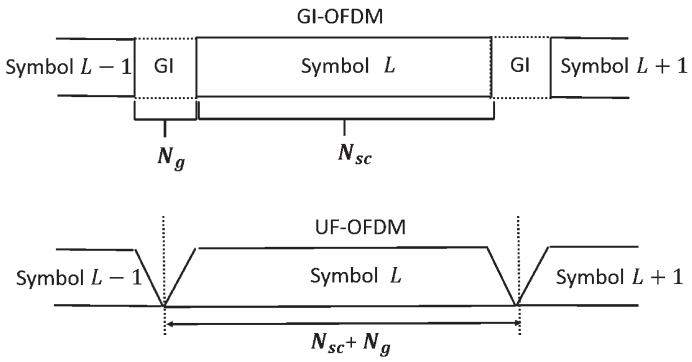


FIGURE 5.13

A guard interval is required for interference-free filtering in UF-OFDM modulator.

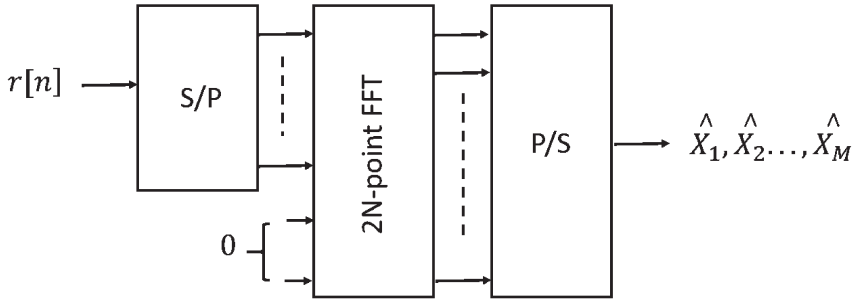


FIGURE 5.14

UF-OFDM receiver structure.

receiver utilizes complete symbol including the GI part in the demodulation. A $2N_{sc}$ point fast Fourier transform (FFT) is performed to recover the data. From the FFT output block (of size $2N_{sc}$), only the even indexed symbols are the desired reconstructed QAM symbols. For details, we refer the reader to [12,11].

5.1.3 FILTER BANK-BASED WAVEFORMS

There is an inverse relation between time and frequency. That is, extending a signal in the time domain makes it shrink in the frequency domain. Based on this principle, the filter bank multicarrier (FBMC) waveforms employ pulse shapes (prototype filters) with a much longer time duration than the OFDM-based waveforms to achieve a more confined spectrum. Spectral confinement is important for efficient utilization of the spectrum, which is a precious resource. However, as the waveform symbols get longer in the time domain, the adjacent waveform symbols overlap. (Non-overlapping symbols would result in loss of data transmission rate.) The FBMC waveforms are not orthogonal. That is, even in the absence of any channel impairment, it is not possible to perfectly reconstruct the QAM symbols transmitted using an FBMC waveform.

A conceptual illustration of the FBMC transmitter is given in Fig. 5.15. We note that conceptually the construction of FBMC is similar to OFDM-based waveforms (e.g., W-OFDM), however, with the

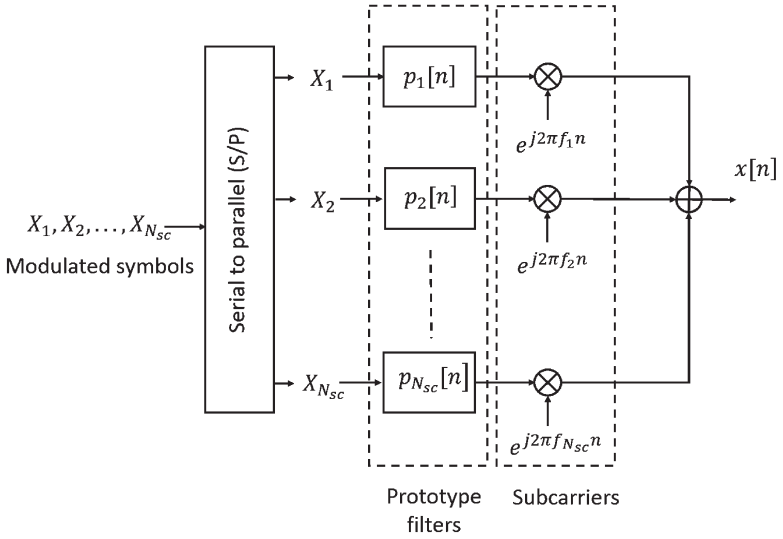


FIGURE 5.15

A conceptual illustration of FBMC modulator.

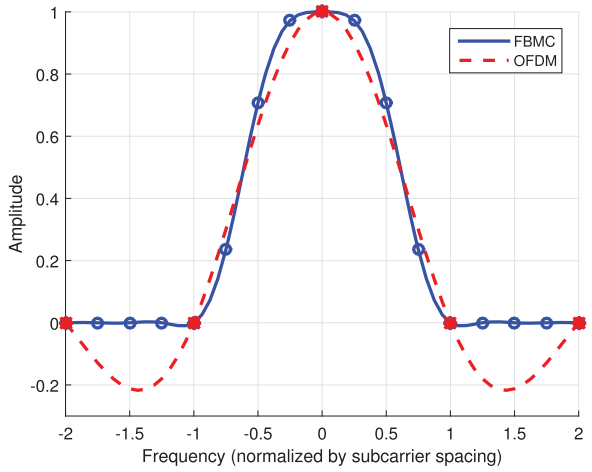


FIGURE 5.16

Frequency responses of the prototype filters of OFDM and FBMC ($K = 4$). The respective frequency domain coefficients are given in Table 5.1.

important difference that the length of the prototype filter is approximately K times greater than the prototype filter (window/pulse) used in the OFDM-based waveforms. For FBMC, the duration of the prototype filter $p_i[n]$ is KN_{sc} samples. The prototype filters for FBMC can be synthesized via frequency domain oversampling of the (rectangular) prototype filter used for OFDM. That is, a prototype filter of length KN_{sc} can be constructed by oversampling the frequency response with a factor K. In Fig. 5.16, we show an example of an oversampled frequency response of the rectangular prototype

Table 5.1 Frequency domain coefficients of the FBMC prototype filters

Waveform	H[0]	H[1]	H[2]	H[3]
OFDM	1	0	0	0
FBMC ($K = 2$)	1	$\sqrt{(2)}/2$	0	0
FBMC ($K = 3$)	1	0.911438	0.411438	0
FBMC ($K = 4$)	1	0.971960	$\sqrt{(2)}/2$	0.235147

The filters have symmetric frequency response, i.e., H[k]=H[-k]

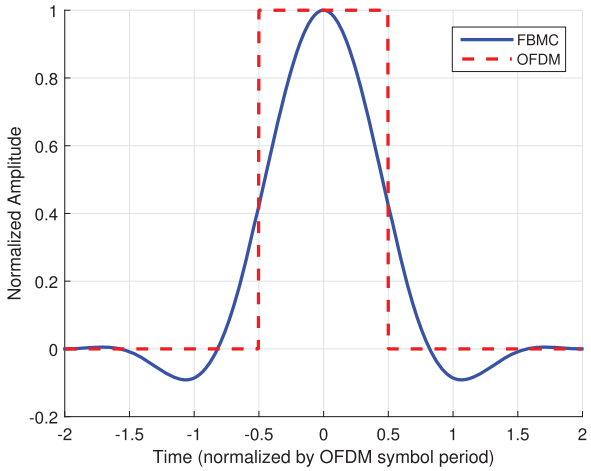


FIGURE 5.17

Impulse responses of the prototype filters used in FBMC ($K = 4$) and OFDM.

filter (used in OFDM) with a factor $K = 4$. The frequency axis in Fig. 5.16 is normalized by $\frac{1}{N_{sc}}$ which is the subcarrier spacing in OFDM and FBMC waveforms. The frequency domain coefficients of the filters are given in Table 5.1. The corresponding time domain representations of these filters are given in Fig. 5.17. The time axis here is normalized by the number of samples in the OFDM symbol without the CP (N_{sc}). The prototype filter is designed to be half-Nyquist, that is, convolution of the transmitter side half-Nyquist and the receiver side half-Nyquist filter is a Nyquist filter that crosses zero (on the time axis) at all the integer multiples of the symbol period [2].

The design, implementation, and characteristics of the prototype filters are discussed in great detail in [2,1]. There are two commonly used methods for implementing FBMC: frequency spreading (FS) and polyphase networks (PPN). In the frequency spreading (FS) method, offset QAM (OQAM) symbols are filtered in the frequency domain, followed by a KN_{sc} point IFFT and an overlap and sum operation. At the receiver side, a sliding window selects KN_{sc} points every $N_{sc}/2$ samples. A KN_{sc} point FFT is performed followed by a matched filtering. In the PPN based FBMC transmitter, an N_{sc} point IFFT is performed on OQAM symbols. The IFFT output (time domain signal) is fed to a polyphase network. At the receiver side, matched filtering is performed followed by a N_{sc} point FFT [2].

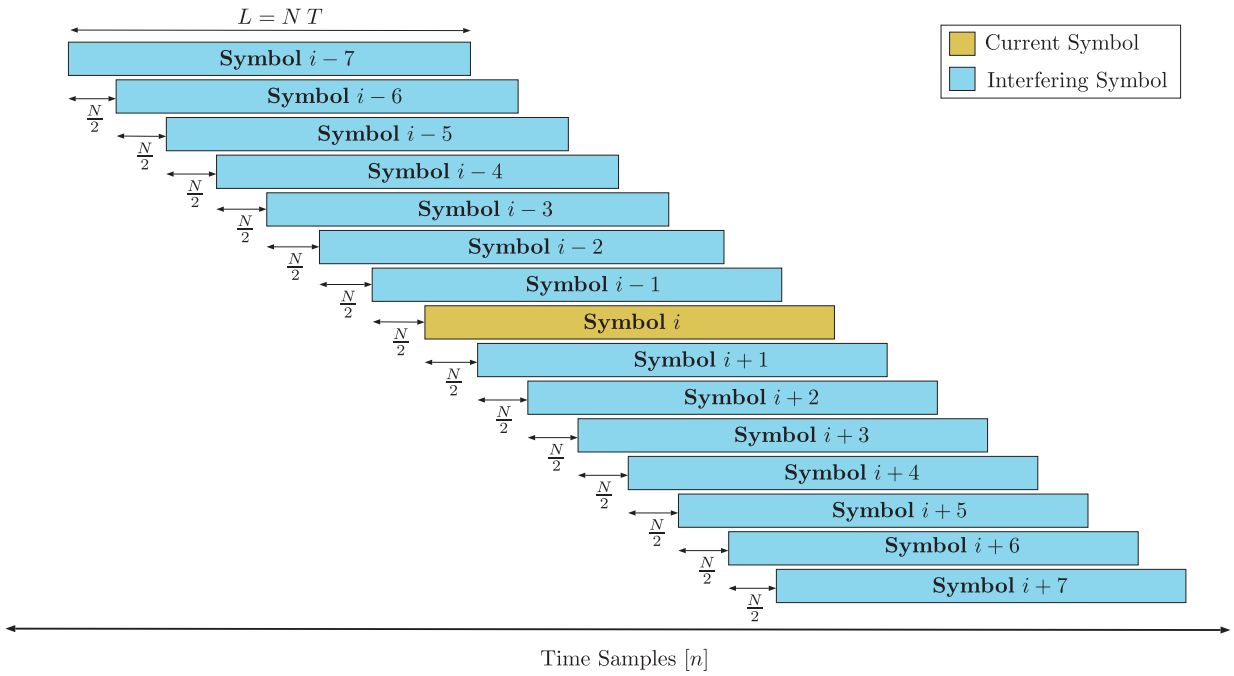


FIGURE 5.18

The overlapping symbol structure of FBMC-OQAM with overlapping factor $K = 4$.

As mentioned earlier, the FBMC waveforms are not orthogonal. There are two reasons for the non-orthogonality:

1. **Intersymbol Interference** An FBMC symbol spans KN_{sc} samples, therefore, the neighboring FBMC symbols overlap as shown in Fig. 5.18. There is a time offset N_{sc} between the adjacent symbols. Due to the overlapping symbols, there is intersymbol interference (ISI) [2]. For a given subcarrier of an FMBC symbol, the ISI primarily comes from the direct adjacent subcarriers of the other overlapping FBMC symbols, i.e., a kth subcarrier in an FBMC symbol primarily suffers interference from subcarriers k and $k-1$ of the overlapping FBMC symbols. The ISI from the subcarriers that are further than the direct neighboring subcarriers is negligible (uniformly upper bounded by -60 dB [2]). There is no ISI between same subcarriers of the overlapping FBMC symbols due to the Nyquist property of the prototype filters [2].
2. **Intercarrier Interference** The subcarriers within an FBMC symbol are not orthogonal either [2]. We refer to this as intercarrier interference (ICI). The ICI primarily comes from the adjacent subcarriers (as in the case of the ISI). For any adjacent subcarriers, there is interference between the real parts or the imaginary parts, but there is no interference between the real and the imaginary parts. To eliminate the ICI between the real parts (or the imaginary parts) of the adjacent subcarriers, they have to be offset by $N_{sc}/2$ samples. This property motivates employing the OQAM modulation, where the subcarriers of the FBMC symbol are modulated with an alternating pattern of real and imaginary valued symbols [2].

Broadly speaking, the FBMC waveforms can be divided into two types—one that employs the offset QAM and one that employs the regular QAM. In the following, we briefly discuss both types.

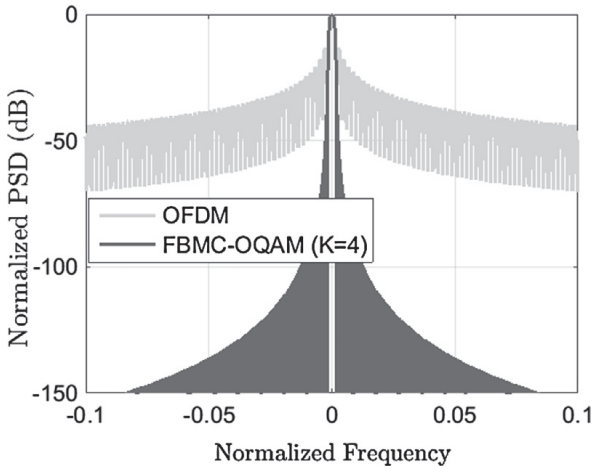


FIGURE 5.19

Comparison of spectrum of OFDM and FBMC-OQAM for one subcarrier.

5.1.3.1 FBMC-OQAM

A discrete-time FBMC-OQAM signal is given by

$$x[n] = \sum_{s\in\mathbb{Z}} \sum_{i\in I} X_{i,s}\, \theta_{i,s}\, p_i \left[n - s\frac{N_{sc}}{2} \right], \tag{5.10}$$

where the set I contains indices of the active subcarriers, N_{sc} is the total number of subcarriers, $X_{i,s} \in \mathbb{R}$ is the modulation symbol of the subcarrier i and the (FBMC) symbol time index s, $\theta_{i,s} := j^{i+s}$ is the phase rotation associated with the offset QAM mapping, $p_i[n] := p[n]\, \mathrm{e}^{j2\pi \frac{i}{N_{sc}} n}$ is the filter for the subcarrier i with $p[n]$ as the base filter with length $L = K\, N_{sc}$ (where K is the overlapping/oversampling factor). The overlapping factor K is a parameter of the FBMC-QAM waveform which refers to the property that a given FBMC-OQAM symbol overlaps with $2K - 1$ symbols preceding it and $2K - 1$ symbols following it. Furthermore, $D = \{-2K + 1, ..., 2K - 1\}$ denotes the set of the adjacent FBMC symbols that overlap in time. An illustration of the overlapping symbol structure is given in Fig. 5.18 with a symbol overlapping (or frequency domain oversampling) factor $K = 4$.

Next, we compare spectral shapes of OFDM and FBMC-OQAM waveforms considering the widely adopted FBMC prototype filters [2,1]. The frequency domain coefficients $H[k]$ of these prototype filters are given in Table 5.1 for three different oversampling (overlapping) factors. The frequency response is symmetric around the center frequency (which is necessary for ensuring the Nyquist property), i.e., $H[-k] = H[k]$. (The prototype filter in Fig. 5.16 and Fig. 5.17 is also constructed using the coefficients given in Table 5.1.) In Figs. 5.19–5.22, we compare power spectral densities of OFDM and FBMC-OQAM waveforms with different overlapping (oversampling) factors. The comparisons include the spectrum of a single subcarrier as well as of a group of 300 subcarriers. These results show that FBMC-OQAM has significantly sharper spectral roll-off than OFDM, especially for $K \geq 3$. The overlapping factor $K = 4$ is most commonly used in the literature.

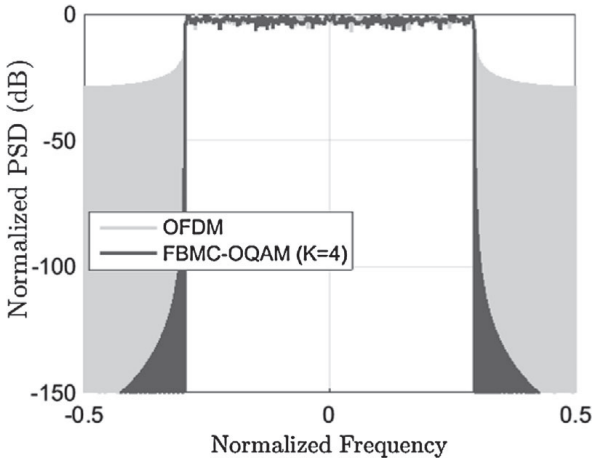


FIGURE 5.20

Comparison of spectrum of OFDM and FBMC-OQAM for 300 subcarriers.

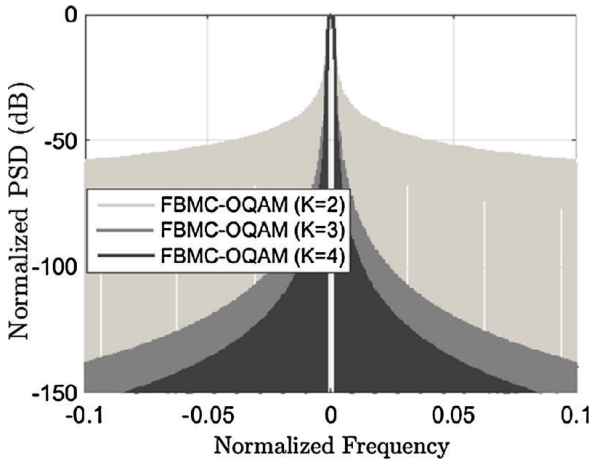


FIGURE 5.21

Comparison of spectrum of one subcarrier of FBMC-OQAM for different overlapping factors.

5.1.3.2 FBMC-QAM

An FBMC waveform that employs a regular QAM (instead of the offset QAM) is referred to as FBMC-QAM. As discussed earlier, there can be significant interference between the real parts (or between the imaginary parts) of the adjacent subcarriers in an FBMC symbol. The offset QAM is one way to reduce the effect of interference from the adjacent subcarriers. Another approach is to design the prototype filters such that the self-interference is minimized or kept to an acceptable level so that the regular QAM can be employed. Typically, the self-interference can be reduced at the cost of degradation in spectral confinement. We will see this shortly in the sequel.

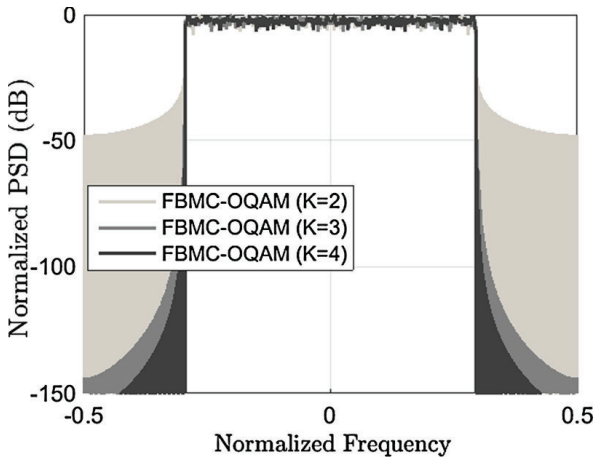


FIGURE 5.22

Comparison of spectrum of 300 subcarriers of FBMC-OQAM for different overlapping factors.

A discrete-time FBMC-QAM signal is expressed as

$$x[n] = \sum_{s\in\mathbb{Z}} \sum_{i\in I} X_{i,s}\; p_i[n - sN_{sc}]\,, \tag{5.11}$$

where I is the set of indices of the active subcarriers, N is the total number of subcarriers, $X_{i,s} \in \mathbb{C}$ is the modulated symbol corresponding to subcarrier i and time index s, and $p_i[n]$ is the filter for subcarrier i. The length of $p_i[n]$ is $L = K\,N$ and K is the overlapping factor. The overlapping factor K is a parameter of FBMC-QAM waveform which refers to the property that a given FBMC symbol overlaps with $K-1$ symbols preceding it and $K-1$ symbols following it. Thus, $D = \{-K+1, ..., K-1\}$ is the set containing indices of the adjacent FBMC symbols that overlap in time considering an overlapping factor K. An example is shown in Fig. 5.23 for a symbol overlapping (or the frequency domain oversampling) factor $K = 4$.

In the recent publications [8,13], a method to design multiple prototype filters (different filters for different subcarriers) has been developed for the synthesis of an FBMC-QAM waveform. To design the prototype filters, a constrained optimization problem is formulated with the objective function being self-interference and the constraints being the spectrum confinement described by falloff rate and the number of nonzero filter taps in the frequency domain. The prototype filters obtained through this method trade-off signal-to-interference ratio caused by the self-interference with spectrum confinement. In [8,13], the authors provide a design example with two prototype filters—one filter is used for the even-numbered subcarrier indices and the other filter is used for the odd-numbered subcarrier indices. That is,

$$p_i[n] := \begin{cases} p_{even}[n]\, \mathrm{e}^{j2\pi \frac{i}{N_{sc}} n} & \text{if } i \text{ is even,} \\ p_{odd}[n]\, \mathrm{e}^{j2\pi \frac{i}{N_{sc}} n} & \text{if } i \text{ is odd.} \end{cases} \tag{5.12}$$

The length of $p_{even}[n]$ and $p_{odd}[n]$ is $L = K\,N_{sc}$. Considering these two prototype filters (see [8,13] for the filter coefficients $p_{even}[n]$ and $p_{odd}[n]$) for FBMC-QAM, we compare its power spectrum

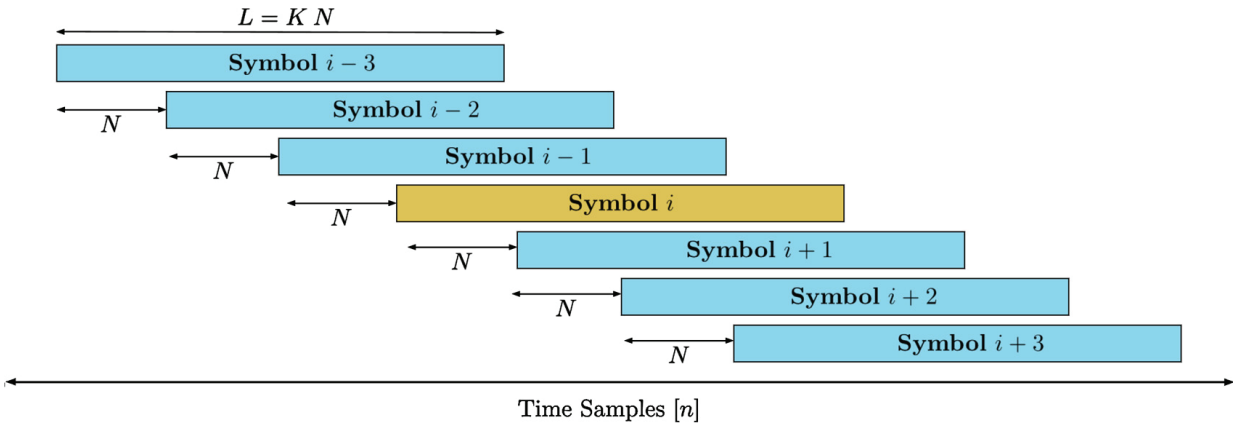


FIGURE 5.23

Overlapping symbol structure of FBMC-QAM with overlapping factor $K = 4$.

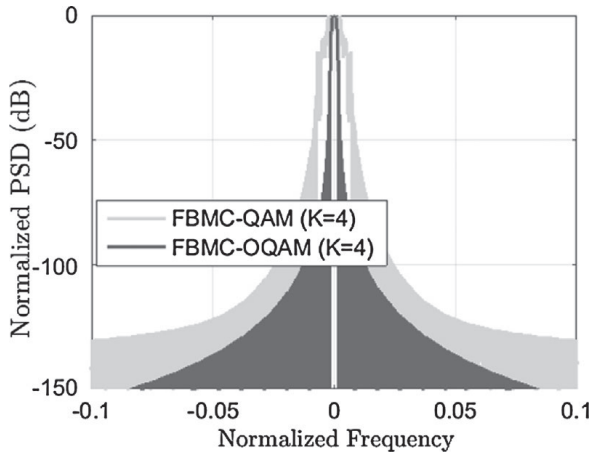


FIGURE 5.24

Comparison of spectrum of FBMC-QAM and FBMC-OQAM for one subcarrier.

with an FBMC-OQAM waveform for a single subcarrier as well for a group of 300 subcarriers in Fig. 5.24 and Fig. 5.25. We observe that the FBMC-OQAM has better spectral characteristics than the FBMC-QAM. A major benefit of FBMC-QAM is that it is better compatible with MIMO transmission due to the utilization of the complex valued QAM symbols.

An important attribute of the FBMC waveforms is that they typically do not employ any CP or GI, unlike the OFDM-based waveforms. The prototype filters in the FBMC waveforms have long decaying tails which can make them robust to the ISI caused by frequency-selective channels. By avoiding the CP overhead, the FBMC waveforms can potentially utilize transmission resources more efficiently than the OFDM-based waveforms, at least in theory. In order to compare the resource usage efficiency, we have sketched the time-frequency lattices for OFDM, FBMC-OQAM, and FBMC-QAM in Fig. 5.26, Fig. 5.27, and Fig. 5.28, respectively. In these figures, Δf denotes the subcarrier spacing

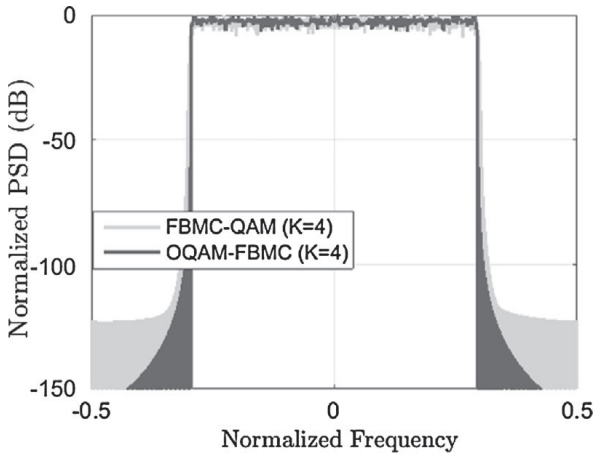


FIGURE 5.25

Comparison of spectrum of FBMC-QAM and FBMC-OQAM for 300 subcarriers.

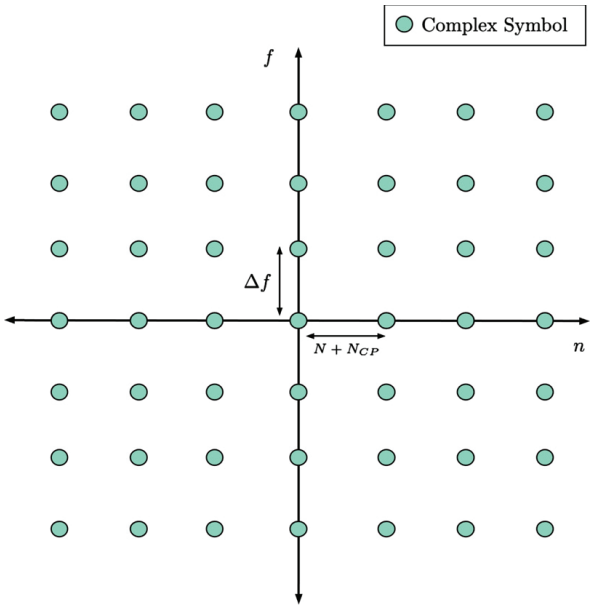


FIGURE 5.26

Time-frequency lattice representation for CP-OFDM.

(which is equal to $\frac{1}{N}$ for all waveforms), N denotes the total number of subcarriers (assumed equal for all waveforms), and N_{CP} is the number of CP samples in CP-OFDM. We observe that FBMC-QAM has one complex symbol per unit area on its lattice and FBMC-OQAM has two real symbols per unit area on its lattice. This means that the two FBMC waveforms have an equal resource usage efficiency. CP-OFDM has $\frac{N}{N+N_{CP}}$ (less than one) complex symbol per unit area, implying a lower resource usage efficiency than the FBMC waveforms. Although FBMC has better time-frequency re-

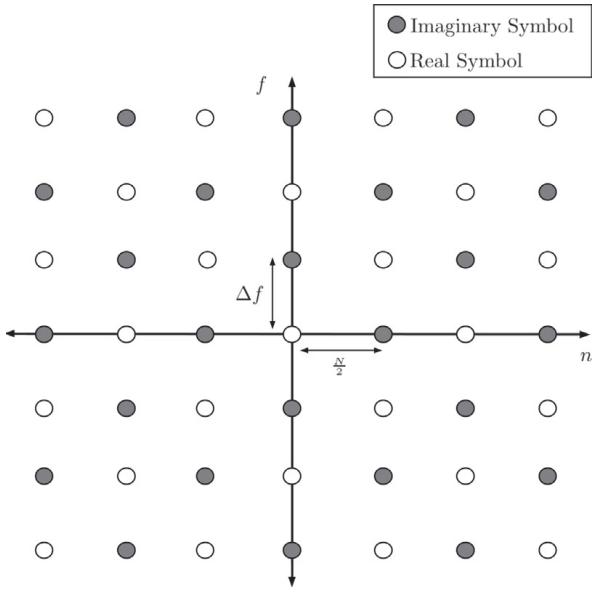


FIGURE 5.27

Time-frequency lattice representation for FBMC-OQAM.

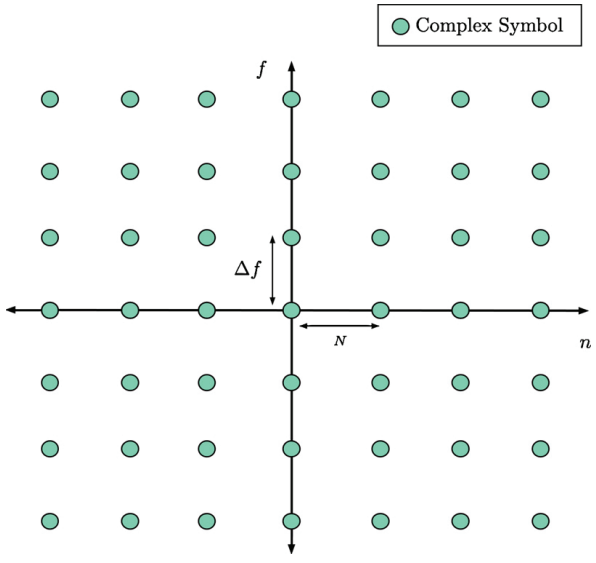


FIGURE 5.28

Time-frequency lattice representation for FBMC-QAM.

source utilization in theory, one cannot simply conclude that CP-OFDM has worse spectral efficiency in a realistic scenario. Here, it is important to keep in mind that the FBMC waveforms suffer from self-interferences (being nonorthogonal waveforms). Moreover, when the channel delay spread is large, the

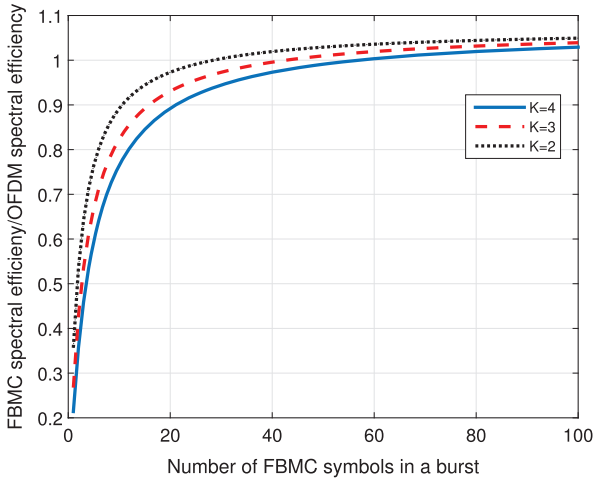


FIGURE 5.29

The ratio of FBMC and OFDM resource usage efficiencies as a function of the number of the waveform symbols per transmission burst.

FBMC waveforms can also have noticeable ISI due to the multipath effect and may require the CP overhead.

Another important aspect is that the time-resource usage efficiency of the FBMC waveforms depends on the transmission burst duration, whereas the time-resource usage efficiency of the OFDM waveforms does not depend on the burst duration. For example, if there are M FBMC symbols in a transmission burst (time slot), then the burst duration has to be $(M + K - 1)N_{sc}$ samples due to the fact that each FBMC symbol has length KN_{sc} and the consecutive FBMC symbols overlap with an offset N_{sc}. This implies that the time-resource usage efficiency of FBMC is $\frac{M}{M+K-1}$. For the OFDM waveform the time-resource utilization efficiency is $\frac{N}{N+N_{CP}}$, regardless of the transmission burst duration. In Fig. 5.29, we plot the ratio of the resource utilization efficiencies of the FBMC and OFDM waveforms. We observe that OFDM is more efficient for shorter transmission burst durations, which is especially important for delay sensitive applications. For LTE, a subframe consists of merely 14 OFDM symbols and therefore OFDM is suitable for LTE like assumptions. In NR, a transmission burst can be as short as one OFDM symbol to support latency critical applications (see Section 2.3 for the mini-slot based transmission).

5.2 SINGLE CARRIER DFTS-OFDM

Multicarrier waveforms, including OFDM, have large variations in instantaneous amplitude (and power) due to the superposition of a large number of modulated subcarriers. When there are large variations in the amplitude of the transmit signal, the power amplifiers of the transmitter either operate in a nonlinear region and cause signal distortion (due to nonlinear clipping) or the transmitter dissipates more power in order to operate in a linear region and prevent distortions. In contrast, single-carrier waveforms have small signal variations and therefore higher power efficiency than multicarrier

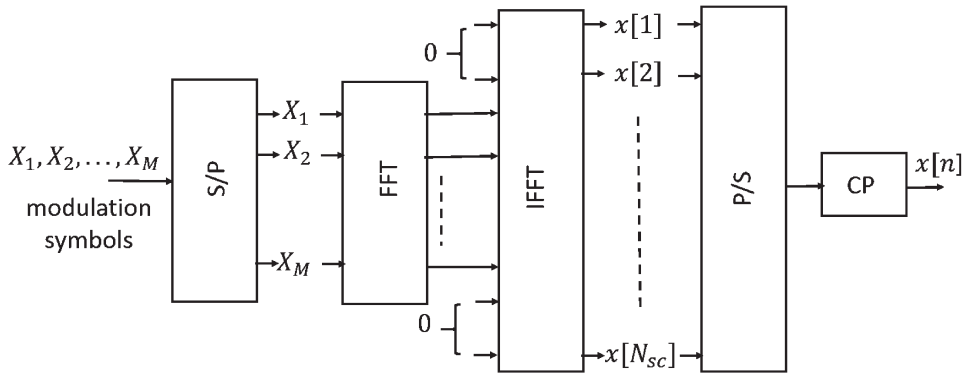


FIGURE 5.30

DFTS-OFDM transmitter structure.

waveforms. For example, there exist constant envelope single-carrier waveforms (e.g., minimum shift keying (MSK), Gaussian minimum shift keying (GMSK)) that allow power amplifiers to run at saturation point without any need for precompensation or postcompensation to account for the clipping. The single carrier PSK modulated waveform also has a constant amplitude. However, pure single-carrier waveforms have their disadvantages—a low spectral efficiency, performance degradation in frequency-selective channels, and lack of flexible resource allocation in the frequency domain. The single-carrier DFT spread OFDM (DFTS-OFDM) waveform aims at combining benefits of the single-carrier and the multicarrier waveforms, with the following key properties:

- Small amplitude variations (single-carrier aspect).
- Flexible resource allocation in the frequency domain (multicarrier aspect).

DFTS-OFDM can be seen as an OFDM waveform with DFT-based precoding (which makes it single-carrier) [7]. The structure of the DFTS-OFDM modulator is shown in Fig. 5.30. An M-point FFT (DFT) is performed on M QAM symbols followed by an N_{sc}-point IFFT with $M < N_{sc}$. A cyclic prefix is inserted as in CP-OFDM to enable low complexity frequency domain equalization at the receiver side. Fig. 5.31 compares the peak-to-average power ratios of CP-OFDM and DFTS-OFDM waveforms via the complementary cumulative distribution function (CCDF) of PAPR, assuming 16 QAM modulation and 1200 subcarriers. CCDF shows the likelihood of PAPR (in terms of percentage) exceeding a given value. We observe a reduced PAPR with DFTS-OFDM. If one employs PSK modulation instead of QAM, PAPR of DFTS-OFDM can be further reduced.

The bandwidth of a DFTS-OFDM signal is given by $\frac{M}{N} f_s$, where f_s is the sampling frequency. The instantaneous signal bandwidth can vary by changing the block size M, to allow for flexible bandwidth assignment. Moreover, users can be multiplexed in the frequency domain by allocating subsets of subcarriers to different users. For example, according to Fig. 5.30, if we want to multiplex two users with equal bandwidths, the first $\frac{M}{2}$ QAM symbols $(X_1, X_2, \ldots, X_{M/2})$ can be associated with user 1 and the last $\frac{M}{2}$ QAM symbols $(X_{M/2}, X_{M/2+1}, \ldots, X_M)$ can be associated with user 2 (assuming M is even). In this example, each user is assigned contiguous subcarriers. Alternatively, each user can also be assigned noncontiguous subcarriers. For example, user 1 and user 2 can be interleaved by assigning odd-numbered subcarriers (i.e., modulation symbols $X_1, X_3, \ldots, X_{M-1}$) to user 1 and even-numbered

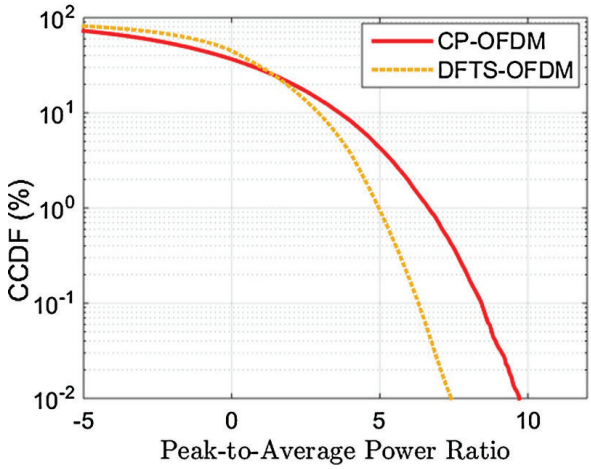


FIGURE 5.31

DFTS-OFDM achieves much reduced PAPR than CP-OFDM.

subcarriers (i.e., modulation symbols $X_2, X_4, \ldots, X_M$) to user 2. When multiple users are multiplexed in DFTS-OFDM waveform, it is often referred to as single-carrier frequency division multiple access (SC-FDMA). If users are assigned contiguous subcarriers, it is called localized frequency division multiple access (LFDMA). When each user is assigned noncontiguous subcarriers that are uniformly distributed across its bandwidth, it is commonly known as interleaved frequency division multiple access (IFDMA).

DFTS-OFDM is used in 4G LTE for the uplink transmissions due to its improved power efficiency. It is also an optional waveform for the 5G NR uplink. However, there are some shortcomings of employing DFTS-OFDM, which are discussed in Chapter 6.

5.3 WAVEFORM DESIGN REQUIREMENTS FOR 5G NR

5G radio access will support various applications (eMBB, URLLC, massive machine type communications (mMTC)), a wide range of frequencies (from sub-1 GHz to 100 GHz), diverse deployments, and various link types (uplink, downlink, device-to-device link, backhaul link). In the following, we provide an overview of important waveform design requirements for different scenarios.

eMBB has high requirements on throughput, user density, and low latency. Both small and large cell deployments are expected for eMBB services. Typically, larger cells in low carrier frequencies in the licensed spectrum and small cells at higher carrier frequencies (e.g., millimeter-wave) in both the licensed and the unlicensed spectrum. For the downlink transmission, a high spectral efficiency and maximum utilization of MIMO are key requirements for the NR waveform. For the uplink transmissions in the large cell deployments, the cell edge users can be power limited and may not be able to achieve high spectral efficiencies. In this case, the power efficiency of the waveform is important. In the small cell deployments, the distance between the users and the base station can significantly reduce

but not necessarily the maximum transmission power of the users. In this case, the power efficiency and the MIMO compatibility are important design requirements.

At very high carrier frequencies (e.g., millimeter-wave), analog or hybrid beam-forming is expected, at least in the early stage of the NR deployments (see Chapter 7 for details). In such systems, the number of digital chains is small compared to the number of antennas and part of the beam-forming is realized via the RF beam-forming. With the RF beam-forming, it is not possible to multiplex users in the frequency domain (frequency division multiplexing (FDM)) unless they all are served by the same beam. In this case, the users should be multiplexed in the time domain (time division multiplexing (TDM)) and therefore a short transmit time interval (TTI) is important. The NR waveform has to be confined in time to allow for the short transmissions.

For the URLLC services, the ultra-reliability and very low latency are important. Beam-forming is an important tool for improving the reliability. To achieve a very low latency, the waveform should be confined in the time domain with a low processing latency. For a fast uplink access, an asynchronous transmission may be beneficial, which means that the waveform should be able to efficiently support the asynchronous access. The retransmission mechanisms are important for enhancing the link reliability. In order to reduce latency due to the HARQ, the waveform should efficiently support the short TTIs.

The mMTC use cases target at providing connectivity to a massive number of devices with strict requirements on battery life and processing complexity. mMTC applications are expected in large cell deployments at lower carrier frequencies (e.g., below 6 GHz). mMTC data traffic can be in the form of short data bursts. Moreover, the radio propagation and penetration losses can be high, for example, to provide connectivity to devices in the basement of a building. Hence, the important waveform design requirements for the mMTC are the high power efficiency (in the uplink), the low complexity transceiver processing, the suitability to the short burst durations, and the ability to multiplex large number of users. Asynchronous access can also be important in the uplink.

The D2D communication typically has short range; therefore, the power efficiency is not very important. For the D2D communication, requirements are expected to be high on throughput and therefore the spectral efficiency is important. Since the D2D links are symmetric in nature, it is desirable to have the same waveforms in both directions.

For the backhaul links (base station-to-base station communication), high throughputs and low latencies are important. The wireless backhaul or the self-backhaul is likely with small cell deployments. Hence, the most important waveform design requirements are high spectral efficiency, compatibility with MIMO, and low latency (similar to eMBB). In wireless backhauling, the links are symmetric in nature and, hence, the same waveform is desirable in both directions.

5.4 KEY PERFORMANCE INDICATOR FOR NR WAVEFORM DESIGN

Key characteristics of 5G NR include large channel bandwidths, extreme data rates, ultra-reliability and low latency requirements, harsh propagation conditions, severe RF impairments, massive number of antennas, small sized base stations, and mainly TDD deployments. Considering these aspects, the mmMAGIC project (a European research project[5]) as well as the 3GPP identified key performance

[5] mmMAGIC was an EU project (funded by H2020 program) that developed radio interface concepts and solutions for above 6 GHz mobile radio communications.

indicators (KPIs) for the NR waveform design and evaluated the candidate waveforms for these KPIs. In the following, we present these waveform design KPIs and their importance in different frequency ranges. The key performance indicators for the NR waveform design include:

- **Spectral efficiency** The spectral efficiency is vital to support extreme requirements on data rate, user connection, and traffic densities. In general, the spectral efficiency is more crucial at lower carrier frequencies than at higher frequencies, since the spectrum is not as precious at higher frequencies due to the availability of potentially much larger channel bandwidths.
- **MIMO compatibility** Multiantenna transmission is a driving technology for NR. With the increase in carrier frequency, the number of antenna elements would increase in the base stations as well as in the devices. The use of various MIMO schemes is essential in providing high spectral efficiencies (by enabling SU-MIMO/MU-MIMO) and greater coverage (via beam-forming). Beam-forming is instrumental in overcoming the high transmission losses at very high frequencies (coverage limited scenarios).
- **Low Peak-to-Average-Power-Ratio (PAPR)** A low PAPR is essential for power efficient transmissions from the devices (for example, uplink, sidelink). A low PAPR becomes even more important at very high frequencies. Since small sized low-cost base stations are envisioned at high frequencies, low PAPR is also important for the downlink transmissions.
- **Robustness to channel time selectivity** The robustness to channel time variations is vital in high vehicular speed scenarios. The high speed scenarios are relevant in large cell deployments. The large cell deployments are not expected at very high frequencies due to harsh propagation conditions (coverage limitation). At very high frequencies, the deployments are expected to come in the form of small cells where mobility is not a major concern. However, certain vehicular services may be enabled at very high frequencies.
- **Robustness to channel frequency selectivity** Typically, wideband wireless channels are strongly frequency selective and robustness to frequency selectivity is fundamental to support high throughput communication over wideband channels.
- **Robustness to phase-noise** Phase-noise depends on quality of the local oscillators at the devices and at the base stations. Typically, phase-noise is high due to the UE (transmitter/receiver), since low phase-noise oscillators can be too expensive and power consuming for the devices. Phase-noise robustness is also important for future low-cost base stations. Basically, any link that involves a device and/or low-cost base station puts a high requirement on the phase-noise robustness of the waveform; especially if the communication takes place at high frequencies since phase-noise typically increases with the carrier frequency (see Chapter 4).
- **Robustness to Power Amplifier Non-linearity** The impact of the nonlinear PA increases as the signal bandwidth increases (see Chapter 4).
- **Transceiver baseband complexity**, involved in encoding and decoding of the information embedded in a waveform. The baseband complexity is always very important for the devices, especially from the receiver perspective. For NR, the complexity is even a major consideration for base stations, since a base station can be a small sized access node (especially at high frequencies) with limited processing capability. At very high frequencies and large bandwidths, the receiver may also have to cope with severe RF impairments. A low baseband complexity is also important for faster processing and enabling low latency applications. Moreover, the baseband complexity becomes increasingly important as the signal bandwidth increases.

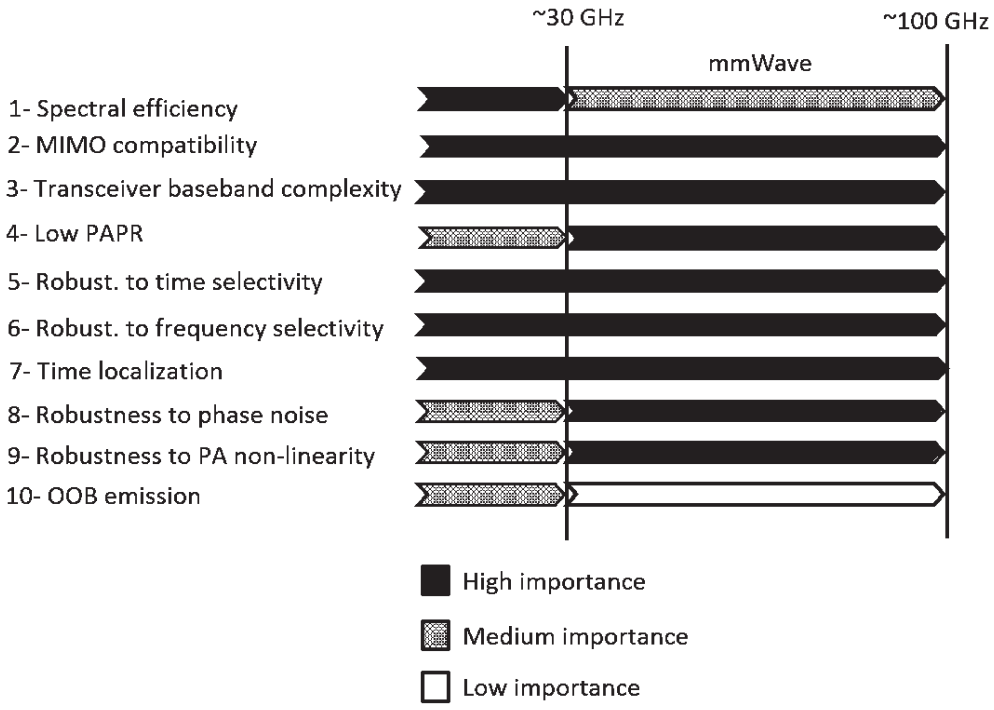


FIGURE 5.32

Importance of waveform performance indicators as a function of carrier frequency.

- **Time localization** Time localization is important to efficiently enable (dynamic) TDD and support low applications. Frequent link direction switching for TDD requires short burst transmissions. A low latency, which is one of the key requirements for both the eMBB and the URLLC services, also requires short transmission time slots. A waveform that is confined in the time domain is suitable for enabling short transmissions.
- **Out-of-band emissions/Frequency localization**: Frequency localization is useful for efficient utilization of spectrum and potential multiplexing of different services (e.g., URLLC, mMTC, eMBB) on a single carrier using different waveform numerologies (see Section 6.3.3). Frequency localization is not a major performance indicator at high frequencies where large channel bandwidths are available. Frequency localization is also relevant for asynchronous access, which can be useful in uplink and sidelink.
- **Flexibility and scalability** Flexibility and scalability of the waveform is important to enable diverse use cases and deployment scenarios.

Fig. 5.32 sketches importance of the waveform KPIs for the NR operation in different frequency ranges. We note that, for millimeter-wave communication, special attention needs to be paid to hardware impairments and power efficiency, whereas frequency localization is not of great importance.

Table 5.2 Comparison summary of multicarrier waveforms. (Source: mmMAGIC project [4,15])

MC-waveforms	CP-OFDM	W-OFDM	UF-OFDM	FBMC-QAM	FBMC-OQAM
Spectral efficiency	High	High	High	High	High
PAPR	High	High	High	High	High
Phase-noise robustness	Medium	Medium	Low/Medium	Low/Medium	Low/Medium
Robust. to freq. selective chan.	High	High	High/Medium	High	High
Robust. to time-selective chan.	Medium	Medium	Open	Low/Medium	Low/Medium
MIMO compatibility	High	High	Open	Open	Low
Time localization	High	High	High	Low	Low
OOB emissions	High	Medium	Medium	Low	Low
OOB emissions with PA	High	High/Medium	High/Medium	High/Medium	High/Medium
Complexity	Low	Low	Medium	High	High
Flexibility	High	High	High	High	High

5.5 WAVEFORM COMPARISON FOR NR

The mmMAGIC project evaluated several multicarrier and single-carrier waveforms based on the waveform design KPIs discussed in the previous section. The project concluded with an overall comparison of the candidate waveforms prior to the NR standardization in the 3GPP. These results served as an important input for the NR standardization. The waveform comparison results are summarized in Table 5.2 for selected multicarrier waveforms. The following color scheme has been used in this table: Green (light gray in print version) refers to a desirable characteristic; red (medium gray in print version) refers to an undesirable characteristic; blue (dark gray in print version) refers to somewhere between desirable and undesirable characteristic. Moreover, "Open", means that further investigations are required to draw any conclusion. The investigations were based on analytical methods as well as simulations. The simulations were performed under common evaluation assumptions in the mmMAGIC waveform simulators. (One of the two simulators is presented in Chapter 9.) The detailed evaluation results are available in the mmMAGIC project report [4] and are summarized in [15]. In the following, we provide the waveform comparison results for selected KPIs (power efficiency, phase-noise robustness, frequency localization, baseband complexity) and refer the reader to [4] for further details.

5.5.1 FREQUENCY LOCALIZATION

CP-OFDM suffers from poor frequency localization/high out-of-band emissions. Several multicarrier waveforms (both OFDM based and FBMC based) aim at improving frequency localization, as shown in Section 5.1.2 and Section 5.1.3. Among these multicarrier waveforms, FBMC-OQAM ($K = 4$) offers the best frequency localization, followed by FBMC-QAM ($K = 4$), and then UF-OFDM and W-OFDM. However, these waveforms behave differently when they are subject to a realistic power amplifier. For example, under the combination of a high transmission power and moderate to large bandwidths, the spectrum roll-off of all waveforms (W-OFDM, UF-OFDM, FBMC) is similar to CP-OFDM, except very close to band edges where FBMC has a steeper roll-off than W-OFDM and UF-OFDM [4].

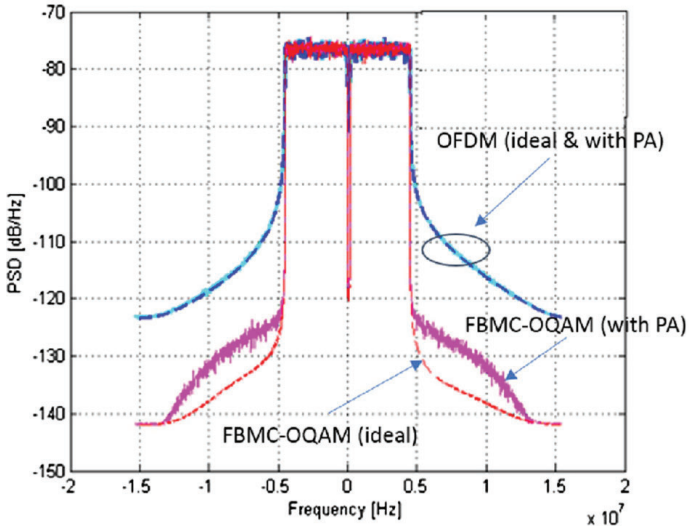


FIGURE 5.33

A PSD comparison of CP-OFDM and FBMC-OQAM with and without PA. Low transmission power and large bandwidth scenario (input power = −21 dBm, output power = 6 dBm, bandwidth = 9 MHz).

In order to show the behavior with respect to nonlinear power amplification, we compare the power spectrum of CP-OFDM and FBMC-OQAM waveforms subject to a power amplifier modeled as a memoryless polynomial in the time domain (as discussed in Chapter 4), according to

$$PA_{output}[n] = PA_{input}[n]p(|PA_{input}[n]|), \tag{5.13}$$

where $p[x] = c_0 + c_1x + c_2x^2 + \cdots + c_kx^k$. For simulations, we have used a tenth order polynomial with coefficients given in [4] and have assumed 2048 subcarriers, 30.72 MHz sampling frequency, and QPSK modulation. For FBMC-OQAM, we have assumed K (overlapping factor) = 4 and the prototype filter given in Table 5.1. In Figs. 5.33–5.35, we compare PSDs of OFDM and FBMC (with and without PA) at different input/output power levels and different signal bandwidths. All PA input/output power values in dBm assume a 50 Ω impedance. We note that the actual transmitter power will be a few (3 to 4) dBs lower due to losses in duplexer/switches. All results are shown without any power back-off. In the following, we summarize the key observations.

- For low transmission power levels or small signal bandwidths, substantially lower OOB emission can be achieved with FBMC compared to OFDM (cf. Fig. 5.33 and Fig. 5.35).
- For high transmission power levels and moderate to high signal bandwidths, the OOB emission is similar for FBMC and OFDM (cf. Fig. 5.35).

In general, for a combination of high transmission power and a large bandwidth, it is unlikely that the sharp spectrum roll-off promised by FBMC and other frequency localized multicarrier waveforms can be realized with state-of-the-art PA technology. In Chapter 8, we provide further evaluations for a power amplifier model with memory.

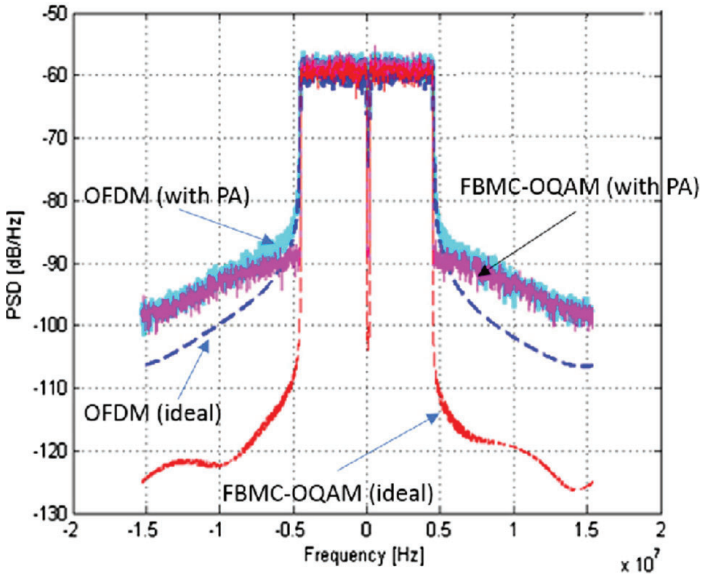


FIGURE 5.34

A PSD comparison of CP-OFDM and FBMC-OQAM with and without PA. High transmission power and large bandwidth scenario (input power = −4 dBm, output power = 24 dBm, bandwidth = 9 MHz).

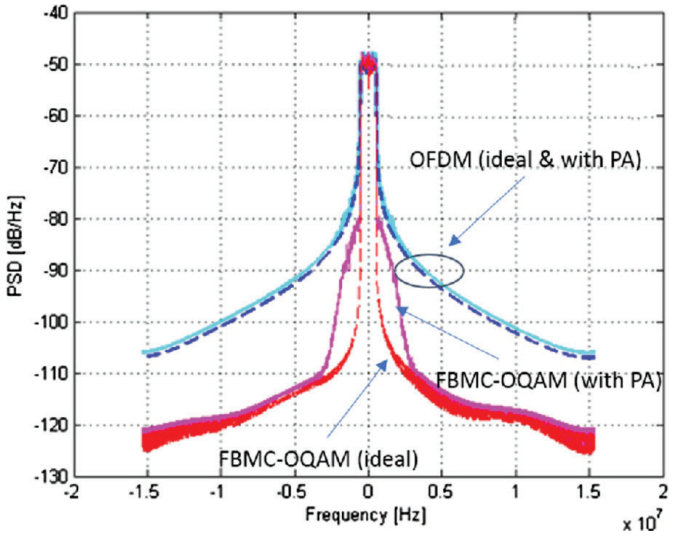


FIGURE 5.35

A PSD comparison of CP-OFDM and FBMC-OQAM with and without PA. High transmission power and small bandwidth scenario (input power = −4 dBm, output power = 24 dBm, bandwidth = 1.08 MHz).

5.5.2 POWER EFFICIENCY

As discussed in Section 5.2, a common drawback of all multicarrier waveforms is their high PAPR (and low power efficiency). In Fig. 5.36, we compare the PAPR of several waveforms, including CP-OFDM,

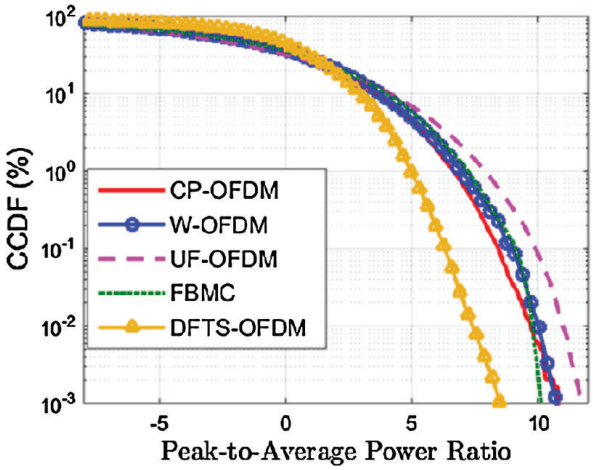


FIGURE 5.36

A comparison of PAPR of multicarrier waveforms and single carrier DFTS-OFDM waveform.

W-OFDM, UF-OFDM, FBMC-OQAM, and DFTS-OFDM (assuming 16 QAM and 1200 subcarriers). We observe that all multicarrier waveforms have similar PAPR except UF-OFDM, which has a higher PAPR. The DFT-based precoding in OFDM (DFTS-OFDM) reduces the PAPR and achieves a higher power efficiency than OFDM. There are various well-known methods to improve the power efficiency of OFDM, which are discussed in Section 6.4.

5.5.3 TIME-VARYING FADING CHANNEL

In Fig. 5.37, we compare performance of several multicarrier waveforms (CP-OFDM, W-OFDM, UF-OFDM, FBMC-OQAM, and FBMC-QAM) in terms of the symbol error rate over a time-varying fading channel with 60 km/h UE speed at 6 GHz carrier frequency (assuming QuaDRiGa channel model). For all waveforms, we assume 16 QAM, 512 subcarriers, and 120 MHz signal bandwidth. As can be seen, CP-OFDM has superior performance in the time-varying fading channel. Further details of this evaluation are given Chapter 9.

5.5.4 BASEBAND COMPLEXITY

We now look at implementation complexity of multicarrier waveforms in terms of the number of real multiplications required for synthesis and demodulation, excluding the computations required for channel estimation. There are different channel estimation techniques used in practice with varying degrees of complexity. We compare complexity of the following multicarrier waveforms: OFDM, W-OFDM, F-OFDM, UF-OFDM, and FBMC. Since FFT/IFFT operations are used in implementations of all of these waveforms, for simplicity we will denote the complexity of an N-point FFT/IFFT operation with $\mathcal{C}_{FFT}(N)$. The results presented in the following are based on [3], where the reader can find a detailed analysis including complexity evaluation for the FPGA-based implementations.

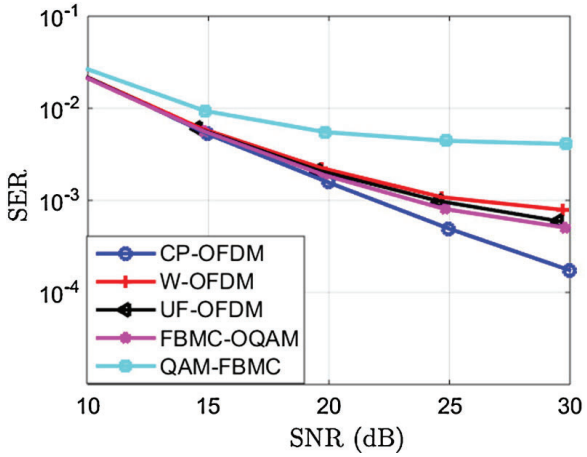


FIGURE 5.37

A comparison of multicarrier waveforms subject to a time-varying fading channel (60 km/h UE speed at 6 GHz carrier frequency).

5.5.4.1 CP-OFDM

Considering the total number of subcarriers N_{sc}, the number of active subcarriers N_{act}, and per subcarrier channel equalization, the number of real multiplications for a CP-OFDM transmitter and receiver are given by

$$\mathcal{C}_{OFDM}^{Tx} = \mathcal{C}_{FFT}(N_{sc}), \tag{5.14}$$

$$\mathcal{C}_{OFDM}^{Rx} = \mathcal{C}_{FFT}(N_{sc}) + 4N_{act}, \tag{5.15}$$

where $4N_{act}$ in (5.15) corresponds to one complex multiplication per subcarrier for channel equalization.

5.5.4.2 W-OFDM

The windowing operation in an OFDM transmitter/receiver is merely multiplication of a window function with an OFDM symbol. Consider an OFDM symbol of length $N_{sc} + N_{cp}$ samples with N_{cp} denoting the number of CP samples, the number of real multiplications for the W-OFDM transmitter and receiver are given by

$$\mathcal{C}_{W-OFDM}^{Tx} = \mathcal{C}_{OFDM}^{Tx} + 4(N_{sc} + N_{cp}) = \mathcal{C}_{FFT}(N_{sc}) + 4(N_{sc} + N_{cp}), \tag{5.16}$$

$$\mathcal{C}_{W-OFDM}^{Rx} = \mathcal{C}_{OFDM}^{Rx} + 4(N_{sc} + N_{cp}) = \mathcal{C}_{FFT}(N_{sc}) + 4N_{act} + 4(N_{sc} + N_{cp}). \tag{5.17}$$

The term $4(N_{sc} + N_{cp})$ corresponds to the number of real multiplications required for the transmitter/receiver windowing operation.

5.5.4.3 UF-OFDM

In the UF-OFDM transmitter, the subcarriers are divided into groups (subbands or resource blocks) and then group-wise filtering is performed. Consider a UF-OFDM symbol with N_{sc} total subcarriers, B groups of N_B subcarriers with group-wise filtering, and a filtering operation in the frequency domain. The number of real multiplications per UF-OFDM symbol is given by

$$\mathcal{C}^{Tx}_{UF-OFDM} = B\left(\mathcal{C}_{FFT}\left(2N_B\right) + \mathcal{C}_{FFT}\left(N_B\right) + 8N_B\right) + \mathcal{C}_{FFT}\left(2N_{sc}\right). \tag{5.18}$$

The UF-OFDM receiver processing includes windowing in the time domain, transformation of the signal to the frequency domain, filtering in the frequency domain, and equalization in the frequency domain. Since the length of a UF-OFDM symbol is $N_{sc} + N_g$ samples (where N_g is the guard interval; cf. Fig. 6.17), the symbol is zero padded and an FFT of size $2N_{sc}$ is performed while we have a transformation from the time to the frequency domain. Considering these operations, the total number of real multiplications is given by

$$\mathcal{C}^{Rx}_{UF-OFDM} = \mathcal{C}_{FFT}(2N_{sc}) + 8N_{sc} + 4N_{act}. \tag{5.19}$$

5.5.4.4 FBMC-OQAM

FBMC waveform can be synthesized via two different approaches: a polyphase network (PPN)-based method and a frequency domain filtering (frequency spread)-based method. The two methods have different implementation complexity. For the PPN implementation, the transmitter and receiver side number of real multiplications (assuming an overlapping factor K for FBMC) are given by

$$\mathcal{C}^{Tx}_{FBMC-PPN} = 2\mathcal{C}_{FFT}(N_{sc}) + 4N_{sc}K + 4N_{act}, \tag{5.20}$$

$$\mathcal{C}^{Rx}_{FBMC-PPN} = 2\mathcal{C}_{FFT}(N_{sc}) + 4N_{sc}K + 4L_{eq}N_{act}, \tag{5.21}$$

where L_{eq} tap channel equalization is performed per subcarrier.

The complexity of the frequency spread FBMC method is given by

$$\mathcal{C}^{Tx}_{FBMC-FS} = 2\mathcal{C}_{FFT}(KN_{sc}) + 8N_{act}(K-1), \tag{5.22}$$

$$\mathcal{C}^{Rx}_{FBMC-FS} = 2\mathcal{C}_{FFT}(KN_{sc}) + 16N_{act}(K-1). \tag{5.23}$$

The transmitter and the receiver side computational complexity of these multicarrier waveforms is compared as a function of the number of subcarriers in Fig. 5.38 and Fig. 5.39, respectively. Here, we have assumed $N_{sc} = N_{act}$ for all waveforms, 7% CP length for OFDM and W-OFDM (i.e., $N_{cp} = .07N_{sc}$), $N_B = 16$ for UF-OFDM, $K = 4$ for FBMC-PPN and FBMC-FS, and $L_{eq} = 3$ for FBMC-PPN. We observe that CP-OFDM has the lowest complexity and the windowing contributes to a minor complexity increase. The complexity of the other multicarrier waveforms is significantly higher than CP-OFDM and the complexity gap enlarges as the number of subcarriers increase. For UF-OFDM, the complexity depends on N_B. As N_B increases, the complexity decreases. We also note that the PPN based FBMC implementation has significantly lower implementation complexity than the frequency spread based implementation.

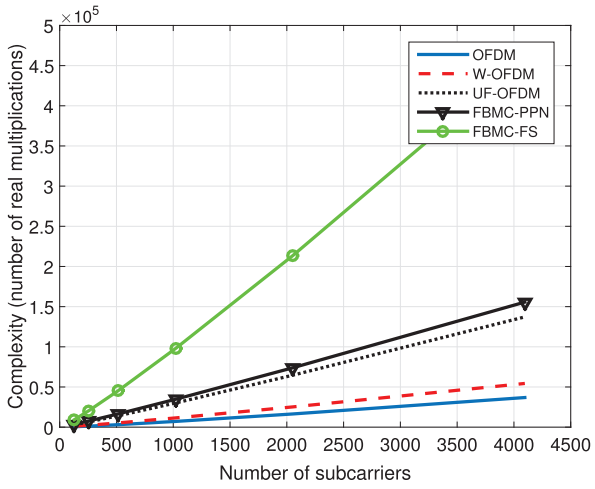


FIGURE 5.38

Transmitter side complexity comparison.

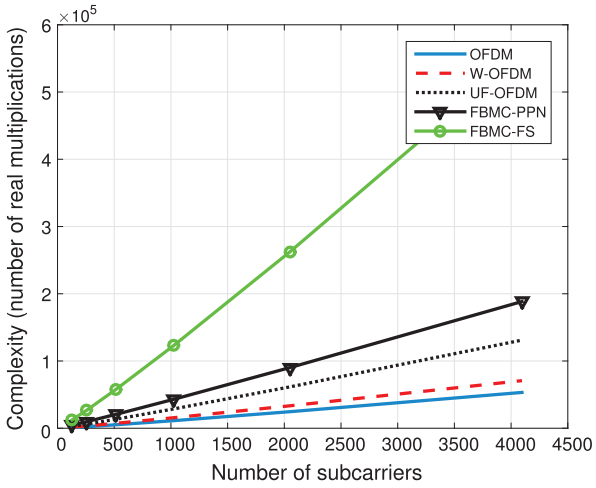


FIGURE 5.39

Receiver side complexity comparison.

5.5.5 PHASE-NOISE ROBUSTNESS COMPARISON

In the following we compare phase-noise robustness of CP-OFDM, FBMC-OQAM, and FBMC-QAM waveforms based on [6]. A baseband signal subject to phase-noise (PN) can be expressed as $r[n] = x[n]\,\mathrm{e}^{j\phi[n]}$, where $\phi[n]$ is a random process. Typically, in a phase-locked loop (PLL)-based oscillators $\phi[n]$ is a wide-sense stationary (WSS) random process, and $|\phi[n]| \ll 1$ [5], which leads to the following approximation:

$$r[n] = x[n]\,\mathrm{e}^{j\phi[n]} \approx x[n](1 + j\phi[n]). \tag{5.24}$$

5.5.5.1 Phase-Noise Effect in OFDM

Assuming that the received OFDM signal $r[n]$ is subject to the PN, consider the demodulation of a subcarrier l in the 0th OFDM symbol (i.e., $s = 0$) without any loss of generality:

$$\begin{aligned} R_{l,0} &= \sum_{n\in\mathbb{Z}} r[n]\,\overline{p_l[n]} \\ &\overset{(a)}{=} \sum_{n\in\mathbb{Z}} \left(\sum_{d\in\mathbb{Z}} \sum_{i\in I} X_{i,d}\; p_i[n-dN] \right) (1+j\phi[n])\,\overline{p_l[n]} \\ &\overset{(b)}{=} \sum_{i\in I} X_{i,0} \sum_{n\in\mathbb{Z}} p_i[n]\,\overline{p_l[n]}(1+j\phi[n]) \\ &\overset{(c)}{=} \sum_{i\in I} X_{i,0}\,(\delta_{i,l}+\varphi_{i,l}) \overset{(d)}{=} X_{l,0}\left(1+\varphi_{l,l}\right) + N_l^{ICI}, \end{aligned} \tag{5.25}$$

where (a) follows from (5.7) and (5.24); (b) follows from the property that consecutive OFDM symbols do not interfere in the time domain ($p_i[n-dN]\,\overline{p_l[n]} = 0\;\forall n \in \mathbb{Z}$ when $d \neq 0$), avoiding intersymbol interference (ISI), (c) follows from $\varphi_{i,l} := \sum_{n\in\mathbb{Z}} j\,\phi[n]\,p_i[n]\,\overline{p_l[n]}$ and $\sum_{n\in\mathbb{Z}} p_i[n]\,\overline{p_l[n]} = \delta_{i,l}$, where $\delta_{i,l}$ is the Kronecker delta function; (d) follows by defining $N_l^{ICI} := \sum_{i\in I_{ICI}} X_{i,0}\,\varphi_{i,l}$, and $I_{ICI} = I \setminus \{l\}$ is the set of subcarriers that interfere with subcarrier l due to the PN. According to (5.25), the transmitted symbol X_l is multiplied by the term $(1+\varphi_{l,l})$. This effect is known as a common phase error (CPE) because it causes an identical phase rotation in all subcarriers (i.e., $\varphi_{l,l} = \varphi_{k,k}$ for any l, k). The additive term N_l^{ICI} refers to an intercarrier interference (ICI), which is different for different subcarriers (that is, $N_l^{ICI} \neq N_k^{ICI}$ for any $l \neq k$).

According to [6], the achievable signal-to-interference ratio (SIR) of the demodulated subcarrier l in the OFDM symbol subject to the PN (in absence of CPE) is given by

$$\mathrm{SIR}_l = \frac{1}{\left(\int_{-0.5}^{+0.5} S_\phi(\nu)\; W_l^{ICI}(\nu)\, d\nu \right)}, \tag{5.26}$$

where

$$W_l^{ICI}(\nu) := \sum_{i\in I_{ICI}} \frac{1}{N^2} \left| \frac{\sin\left(\pi\left(\nu - \frac{i-l}{N}\right)N\right)}{\sin\left(\pi\left(\nu - \frac{i-l}{N}\right)\right)} \right|^2. \tag{5.27}$$

Next, we evaluate the achievable SIR using (5.26) for the PLL based phase-noise model described in Section 4.2.3. The PSDs of the PN model are illustrated in Fig. 5.40 at three oscillator different frequencies: 6 GHz, 28 GHz, and 82 GHz. The achievable SIR as a function of the subcarrier spacing is shown in Fig. 5.41. We observe that the SIR is an increasing function of the subcarrier spacing. This implies that the OFDM system would be more robustness against the PN if a larger subcarrier spacing is used.

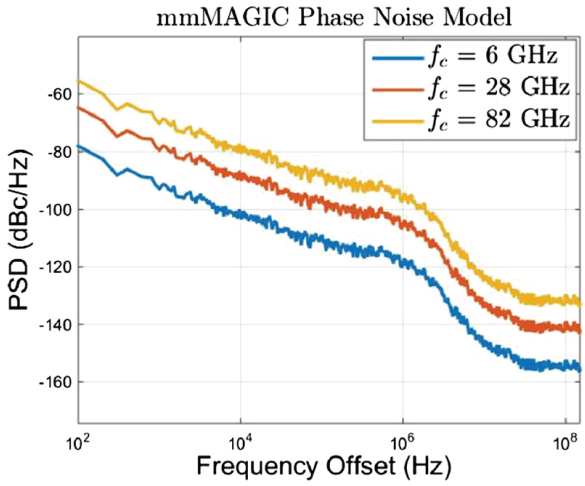


FIGURE 5.40

Power spectrum of the PLL based PN model at three oscillator frequencies: 6 GHz, 28 GHz, and 82 GHz. The model is developed in mmMAGIC (a European research project).

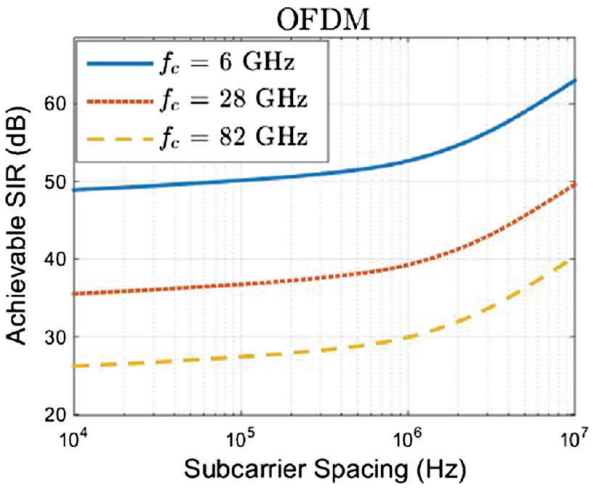


FIGURE 5.41

Achievable signal-to-interference ratio (SIR) for OFDM as a function of its subcarrier spacing at three oscillator frequencies (6 GHz, 28 GHz, and 82 GHz).

5.5.5.2 Phase-Noise Effect in FBMC-QAM

The demodulation of a subcarrier l in the 0th FBMC-QAM symbol (i.e., $s = 0$) subject to the PN is given by

$$R_{l,0} = \sum_{n \in \mathbb{Z}} r[n] \, \overline{p_l[n]}$$

$$\begin{aligned}
&\overset{(a)}{=} \sum_{n\in\mathbb{Z}} \left(\sum_{d\in D} \sum_{i\in I} X_{i,d}\, p_i[n-dN] \right) (1+j\phi[n])\, \overline{p_l[n]} \\
&= \sum_{d\in D} \sum_{i\in I} X_{i,d} \sum_{n\in\mathbb{Z}} p_i[n-dN]\, \overline{p_l[n]}\, (1+j\phi[n]) \\
&\overset{(b)}{=} \sum_{d\in D} \sum_{i\in I} X_{i,d} \left(\lambda_{i,l,d} + \beta_{i,l,d} \right),
\end{aligned} \tag{5.28}$$

where (a) follows from (5.24) and (5.11) and (b) follows by defining

$$\lambda_{i,l,d} := \sum_{n\in\mathbb{Z}} p_i[n-dN]\, \overline{p_l[n]}\,, \tag{5.29}$$

$$\beta_{i,l,d} := \sum_{n\in\mathbb{Z}} j\phi[n]\, p_i[n-dN]\, \overline{p_l[n]}\,. \tag{5.30}$$

The term $\lambda_{i,l,d}$ is due to the self-interference components caused by the overlapping FBMC symbols and the nonorthogonality between the subcarriers, whereas $\beta_{i,l,d}$ is due to the PN; it depends on both the choice of the prototype filters and the PN. We can rewrite (5.28) as

$$R_{l,0} = X_{l,0} \left(1+\beta_{l,l,0}\right) + I_l^{ICI} + I_l^{ISI} + N_l^{ICI} + N_l^{ISI}, \tag{5.31}$$

where we have used the following notations:

$$I_l^{ICI} := \sum_{i\in I_{ICI}} X_{i,0}\, \lambda_{i,l,0}\,, \tag{5.32}$$

$$I_l^{ISI} := \sum_{d\in D_{ISI}} \sum_{i\in I} X_{i,d}\, \lambda_{i,l,d}\,, \tag{5.33}$$

$$N_l^{ICI} := \sum_{i\in I_{ICI}} X_{i,0}\, \beta_{i,l,0}\,, \tag{5.34}$$

$$N_l^{ISI} := \sum_{d\in D_{ISI}} \sum_{i\in I} X_{i,d}\, \beta_{i,l,d}\,, \tag{5.35}$$

where $D_{ISI} = D \setminus \{0\}$ is a set containing the indices of the FBMC-QAM symbols that overlap with the 0th FBMC-QAM symbol. The multiplicative term $(1+\beta_{l,l,0})$ corresponds to the CPE. The CPE has the same value for all even subcarriers and the same value for all odd subcarriers (i.e., $\beta_{k,k,0} = \beta_{l,l,0}$ if $k \bmod 2 = l \bmod 2$ for any l, k). The terms I_l^{ICI} and I_l^{ISI} model the ICI and the ISI due to the self-interferences and the terms N_l^{ICI} and N_l^{ISI} represent the ICI and the ISI due to the PN.

According to [6], the achievable SIR of the demodulated subcarrier l in the FBMC-QAM symbol subject to the PN (in absence of CPE) is given by

$$\mathrm{SIR}_l = \frac{P_X}{P_{I_l^{ICI}} + P_{I_l^{ISI}} + P_{N_l^{ICI}} + P_{N_l^{ISI}}}, \tag{5.36}$$

$$P_{N_l^{ICI}} := P_X \left(\int_{-0.5}^{+0.5} S_\phi(\nu)\, W_l^{ICI}(\nu)\, d\nu \right), \tag{5.37}$$

$$P_{N_l^{ISI}} := P_X \left(\int_{-0.5}^{+0.5} S_\phi(\nu)\, W_l^{ISI}(\nu)\, d\nu \right), \tag{5.38}$$

$$P_{I_l^{ICI}} := P_X\, \delta(\nu)\, W_l^{S-ICI}(\nu), \tag{5.39}$$

$$P_{I_l^{ISI}} := P_X\, \delta(\nu)\, W_l^{S-ISI}(\nu), \tag{5.40}$$

$$W_l^{ICI}(\nu) := \sum_{i \in I_{ICI}} W_{i,l,0}(\nu), \tag{5.41}$$

$$W_l^{ISI}(\nu) := \sum_{d \in D_{ISI}} \sum_{i \in I} W_{i,l,d}(\nu), \tag{5.42}$$

$$W_l^{S-ICI}(\nu) := \sum_{i \in I_{ICI}} W_{i,l,0}^{S}(\nu), \tag{5.43}$$

$$W_l^{S-ISI}(\nu) := \sum_{d \in D_{ISI}} \sum_{i \in I} W_{i,l,d}^{S}(\nu), \tag{5.44}$$

$$W_{i,l,d}(\nu) := W_{i,l,d}^{S}(\nu) = \left| P_i(\nu)\, e^{-j2\pi\nu dN} \circledast \overline{P_l(\nu)} \right|^2, \tag{5.45}$$

where $P_i(\nu) := \mathcal{F}\{p_i[n]\}$, $p_i[n]$ is defined in (5.12), and $\mathcal{F}\{\cdot\}$ is the discrete-time Fourier transform operation.

5.5.5.3 Phase-Noise Effect in FBMC-OQAM

Assuming that the received FBMC-OQAM signal $r[n]$ is subject to the PN, the demodulation of a subcarrier l in the 0th FBMC-OQAM symbol is given by:

$$\begin{aligned}
R_{l,0} &\overset{(a)}{=} \sum_{n \in \mathbb{Z}} \Re r[n]\, \psi_l\, \overline{p_l[n]} \\
&\overset{(b)}{=} \sum_{n \in \mathbb{Z}} \Re \left\{ \left(\sum_{d \in D} \sum_{i \in I} X_{i,d} \theta_{i,d}\, p_i \left[n - d\frac{N}{2} \right] \right) (1 + j\phi[n])\, \psi_l\, \overline{p_l[n]} \right\} \\
&= \sum_{d \in D} \sum_{i \in I} X_{i,d} \sum_{n \in \mathbb{Z}} \Re \left\{ \theta_{i,d}\, p_i \left[n - d\frac{N}{2} \right] \psi_l\, \overline{p_l[n]}\, (1 + j\phi[n]) \right\} \\
&\overset{(c)}{=} \sum_{d \in D} \sum_{i \in I} X_{i,d} \left(\gamma_{i,l,d} + \rho_{i,l,d} \right)
\end{aligned}$$

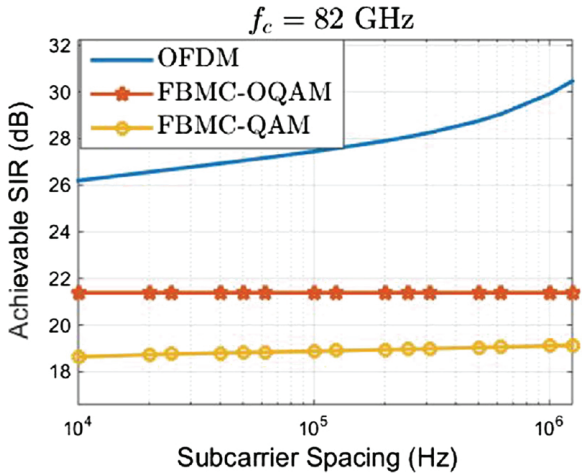


FIGURE 5.42

Achievable signal-to-interference ratio (SIR) for OFDM, FBMC-OQAM ($K = 4$) and FBMC-QAM ($K = 4$) as a function of subcarrier spacing at 82 GHz oscillator frequency.

$$\stackrel{(d)}{=} X_{l,0} + I_l^{ICI} + I_l^{ISI} + N_l^{ICI} + N_l^{ISI}, \tag{5.46}$$

where (a) follows from the offset-demapping which uses the real value operator $\Re\{\cdot\}$ and the phase term $\psi_l = j^{(-l)}$; (b) follows from (5.24) and (5.10); (c) follows by defining

$$\gamma_{i,l,d} := \sum_{n\in\mathbb{Z}} \Re\left\{\theta_{i,d}\; p_i\left[n - d\frac{N}{2}\right]\psi_l\;\overline{p_l[n]}\right\}, \tag{5.47}$$

$$\rho_{i,l,d} := \sum_{n\in\mathbb{Z}} \Re\left\{j\phi[n]\,\theta_{i,d}\; p_i\left[n - d\frac{N}{2}\right]\psi_l\;\overline{p_l[n]}\right\}, \tag{5.48}$$

which are associated with the self-interferences and the PN, respectively; (d) follows from defining I_l^{ICI}, I_l^{ISI} (the same as in (5.32) and (5.33) but replacing $\lambda_{i,l,d}$ by $\gamma_{i,l,d}$), N_l^{ICI} and N_l^{ISI} (the same as in (5.34) and (5.35), but replacing $\beta_{i,l,d}$ by $\rho_{i,l,d}$). It is noteworthy that FBMC-OQAM is not affected by any CPE effect subject to the given PN formulation (i.e., $\forall l \in I\ \rho_{l,l,0} = 0$).

According to [6], the achievable SIR of the demodulated subcarrier l in the FBMC-OQAM symbol subject to the PN is given by (5.36) with the following differences:

$$W_{i,l,d}(\nu) := \frac{1}{4}\left[\left|V_d\left(\nu - \frac{i-l}{N}\right)\right|^2 + \left|V_d\left(\nu + \frac{i-l}{N}\right)\right|^2 - \right.$$
$$\left. - 2\,\Re Y_{i,l,d}(\nu)\right] \tag{5.49}$$

$$W^S_{i,l,d}(\nu) := \frac{1}{4}\left[\left|V_d\left(\nu - \frac{i-l}{N}\right)\right|^2 + \left|V_d\left(\nu + \frac{i-l}{N}\right)\right|^2 + \right.$$
$$\left. + 2\,\Re Y_{i,l,d}(\nu)\right], \tag{5.50}$$

where $Y_{i,l,d}(\nu) := e^{j\pi(i-l+d)} V_d\left(\nu - \frac{i-l}{N}\right) V_d^*\left(\nu + \frac{i-l}{N}\right)$, $V_d(\nu) := P(\nu) \circledast \left(P(\nu)e^{-2\pi\nu\frac{N}{2}d}\right)$, and $P(\nu) := \mathcal{F}\{p[n]\}$.

In Fig. 5.42, we compare the achievable SIRs for OFDM, FBMC-OQAM ($K = 4$), and FBMC-QAM ($K = 4$) as functions of the subcarrier spacing at 82 GHz oscillator frequency. We observe that the FBMC waveforms are more sensitive to phase noise than OFDM.

REFERENCES

[1] M. Bellanger, FS-FBMC: an alternative scheme for filter bank based multicarrier transmission, in: 5th International Symposium on Communications Control and Signal Processing (ISCCSP), 2012, May, pp. 1–4.

[2] M. Bellanger, et al., FBMC physical layer: a primer, http://www.ict-phydyas.org, 2010, 06.

[3] R. Gerzaguet, N. Bartzoudis, L.G. Baltar, V. Berg, J.B. Doré, D. Kténas, O. Font-Bach, X. Mestre, M. Payaró, M. Färber, K. Roth, The 5G candidate waveform race: a comparison of complexity and performance, EURASIP Journal on Wireless Communications and Networking 2017 (1) (2017, Jan.).

[4] J. Luo, A.A. Zaidi, J. Vihriälä, D. Giustiniano, et al., Preliminary radio interface concepts for mm-wave mobile communications. Deliverable D4.1, Millimetre-Wave Based Mobile Radio Access Network for Fifth Generation Integrated Communications (mmMAGIC), https://5g-mmmagic.eu/results, 2016.

[5] A. Mehrotra, Noise analysis of phase-locked loops, in: IEEE/ACM International Conference on Computer Aided Design. ICCAD – 2000. IEEE/ACM Digest of Technical Papers (Cat. No.00CH37140), 2000, Nov., pp. 277–282.

[6] V. Moles-Cases, A.A. Zaidi, X. Chen, T.J. Oechtering, R. Baldemair, A comparison of OFDM, QAM-FBMC, and OQAM-FBMC waveforms subject to phase noise, in: 2017 IEEE International Conference on Communications (ICC), 2017, May, pp. 1–6.

[7] H.G. Myung, J. Lim, D.J. Goodman, Single carrier fdma for uplink wireless transmission, IEEE Vehicular Technology Magazine (ISSN 1556-6072) 1 (3) (2006, Sept.) 30–38, https://doi.org/10.1109/MVT.2006.307304.

[8] H. Nam, M. Choi, C. Kim, D. Hong, S. Choi, A new filter-bank multicarrier system for QAM signal transmission and reception, in: Proc. IEEE ICC, 2014, June, pp. 5227–5232.

[9] F. Schaich, T. Wild, Y. Chen, Waveform contenders for 5G—suitability for short packet and low latency transmissions, in: IEEE Vehicular Technology Conference (VTC Spring), 2014, May, pp. 1–5.

[10] F. Sjoberg, R. Nilsson, M. Isaksson, P. Odling, P.O. Borjesson, Asynchronous zipper, in: IEEE International Conference on Communications (ICC), 1999, pp. 231–235.

[11] T. Wild, F. Schaich, A reduced complexity transmitter for UF-OFDM, in: Proc. IEEE 81st Vehicular Technology Conference (VTC Spring), 2015, May, pp. 1–6.

[12] T. Wild, F. Schaich, Y. Chen, 5G air interface design based on universal filtered (UF-)OFDM, in: 19th International Conference on Digital Signal Processing (DSP), 2014, Aug., pp. 699–704.

[13] Y.H. Yun, C. Kim, K. Kim, Z. Ho, B. Lee, J.Y. Seol, A new waveform enabling enhanced QAM-FBMC systems, in: Proc. IEEE SPAWC, 2015, June, pp. 116–120.

[14] A.A. Zaidi, R. Baldemair, H. Tullberg, H. Bjorkegren, L. Sundstrom, J. Medbo, C. Kilinc, I.D. Silva, Waveform and numerology to support 5G services and requirements, IEEE Communications Magazine (ISSN 0163-6804) 54 (11) (2016, November) 90–98, https://doi.org/10.1109/MCOM.2016.1600336CM.

[15] A.A. Zaidi, J. Luo, R. Gerzaguet, H. Wang, X. Chen, Y. Qi, N. Cassiau, A. Wolfgang, J. Vihriala, A. Kakkavas, T. Svensson, J. Mohammadi, R.J. Weiler, M. Dieudonne, H. Halbauer, V. Moles-Cases, H. Miao, Evaluation of waveforms for mobile radio communications above 6 GHz, in: 2016 IEEE Globecom Workshops (GC Wkshps), 2016, Dec., pp. 1–6.

[16] L. Zhang, A. Ijaz, P. Xiao, M.M. Molu, R. Tafazolli, Filtered OFDM systems, algorithms, and performance analysis for 5G and beyond, IEEE Transactions on Communications (ISSN 0090-6778) 66 (3) (2018, March) 1205–1218, https://doi.org/10.1109/TCOMM.2017.2771242.

CHAPTER

NR WAVEFORM

6

CP-OFDM waveform has been widely adopted for wireless and wireline communication as well as for digital audio and video broadcasting. The Asymmetric Digital Subscriber Line (ADSL) was the first wireline technology that used OFDM for high speed transmission. Digital Audio Broadcasting (DAB) and Digital Video Broadcasting (DVB-T) standards employ OFDM for broadcasting digital audio and video services. Various wireless IEEE and 3GPP standards have adopted OFDM, for example, IEEE 802.11a/WLAN, IEEE 802.16/WiMAX, 3GPP 4G LTE, and more recently 3GPP 5G NR. In this chapter, we discuss the OFDM design for 5G NR.

This chapter is organized as follows. Section 6.1 discusses suitability of CP-OFDM for NR. Section 6.2 describes a scalable OFDM design for NR, as adopted by the 3GPP. Section 6.3 provides a detailed discussion of how the scalable parameters of OFDM should be chosen considering various factors such as quality of service requirements, type of deployment, carrier frequency, user mobility, hardware impairments, and implementation aspects. Section 6.4 discusses the need for improving power efficiency of the NR waveform for coverage limited scenarios, along with a brief overview of the commonly used techniques. The effects of different synchronization errors in a CP-OFDM system are discussed in Section 6.5. Finally, some mitigation techniques for the hardware impairments (phase noise, carrier-frequency offset, and sampling frequency offset) are presented in Section 6.6.

6.1 SUITABILITY OF OFDM FOR NR

For the NR waveform, the design requirements vary depending on the carrier-frequency range and the link type, as discussed in Chapter 5. The 3GPP has selected CP-OFDM waveform for NR after a thorough investigation of several multicarrier and single-carrier waveforms considering requirements that are important for the NR design. In Chapter 5, we have also provided a comparison of state-of-the art waveforms. In the following, we discuss the suitability of OFDM for NR for various link types (uplink, downlink, sidelink,[1] vehicle-to-anything (V2X)[2] link, backhaul link) considering the waveform key performance indicators:

- ***Spectral efficiency*** OFDM is well known for its high spectral efficiency. The spectral efficiency is vital to meet extreme data rate requirements with NR. In general, the spectral efficiency is more crucial at lower carrier frequencies than at higher frequencies due to potentially much larger channel bandwidths at higher carrier frequencies. The spectral efficiency is important for both uplink and

[1] Sidelink refers to direct device-to-device (D2D) communication.

[2] Vehicle-to-anything (V2X) communication includes vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), vehicle-to-network (V2N), and vehicle-to-pedestrian (V2P).

5G Physical Layer. https://doi.org/10.1016/B978-0-12-814578-4.00011-4
Copyright © 2018 Elsevier Ltd. All rights reserved.

the downlink. The requirements are even more stringent for the backhaul link due to the massive volumes of data transmission between the base stations. Vehicular communication can also create capacity bottlenecks in dense urban scenarios with large number of vehicles periodically broadcasting signals in an asynchronous fashion.

- ***MIMO compatibility*** OFDM enables a straightforward use of MIMO technology (see Chapter 7). With increase in the carrier frequency with NR, the number of antenna elements would increase in the base stations as well as in the devices. The use of various MIMO schemes is essential for enhancing spectral efficiency by enabling SU-MIMO/MU-MIMO and achieving greater coverage via beam-forming (see Chapter 7). Beam-forming is instrumental in overcoming high propagation losses at very high frequencies where coverage is limited.
- ***Peak-to-Average-Power-Ratio (PAPR)*** OFDM has a high PAPR like other multicarrier waveforms (see Section 5.5.2). A low PAPR is essential for power-efficient transmissions from the devices (for example, in the uplink and the sidelink). A low PAPR becomes even more important at very high frequencies where coverage can be limited. It is noteworthy that small sized low cost base stations are envisioned at high frequencies, therefore, low PAPR is also important for the downlink transmissions at high carrier frequencies. A high PAPR in OFDM can also be substantially reduced via various well-known PAPR reduction techniques with only minor compromise in performance [10] (see Section 6.4). For NR, OFDM with PAPR reduction is an attractive option for the uplink and the sidelink.

  LTE uses DFTS-OFDM for both uplink and sidelink due to its lower PAPR. However, DFTS-OFDM has certain drawbacks in comparison with OFDM such as lesser flexibility for scheduling and more complex MIMO receiver with degraded link level and system level performance [21]. For NR, DFTS-OFDM is optional for the uplink and can be used only for a single-stream (without MIMO) transmission. Since MIMO transmission is a key component of NR, DFTS-OFDM is not a preferred option for the uplink and the sidelink in general. The use of one waveform (i.e., OFDM) for all link types also makes transceiver designs and implementations symmetric for all transmissions.
- ***Robustness to channel time selectivity*** is vital in high mobility scenarios. High-speed UEs are relevant in large cell deployments. The large cell deployments are not expected at very high frequencies due to harsh propagation conditions (coverage limitation). At very high frequencies, the deployments are expected in the form of small cells where mobility is not a major concern. However, the V2V services may be enabled at very high frequencies, making robustness to the channel time selectivity a very important performance indicator at very high frequencies. Traditionally, the backhaul link is fixed and mobility is not a concern, however, for the envisioned mobile backhaul (e.g., access nodes on vehicles), robustness to channel time selectivity will become relevant. OFDM can be made robust to channel time selectivity by a proper choice of the subcarrier spacing.
- ***Robustness to channel frequency selectivity*** Channel frequency selectivity is always relevant to the transmission of large bandwidth signals over wireless multipath channels. Channel frequency selectivity depends on various factors such as the type of deployment, the beam-forming technique, and the signal bandwidth. OFDM is robust to frequency-selective channels and requires only single-tap frequency-domain equalization.
- ***Robustness against phase noise*** Typically, phase-noise increases with an increase in the oscillator frequency (see Section 4.2). An OFDM system can be made robust to phase-noise by a proper choice of its subcarrier spacing (see Section 6.3.1). Phase-noise robustness is crucial for all link types where a UE (transmitter/receiver) is involved. In particular, low phase-noise oscillators may be

too expensive and power consuming for devices (UEs). Phase-noise robustness is also important for future low cost base stations. Basically, any link that involves a device and/or a low cost base station puts a high requirement on phase-noise robustness of the waveform, especially if the communication takes place at high carrier frequencies.

- ***Transceiver baseband complexity*** The baseband complexity of an OFDM receiver is lowest among all 5G candidate waveforms as discussed in Chapter 5 (see Section 5.5.4 for a complexity comparison). The baseband complexity is always very important at the device (UE) side, especially from the receiver perspective. For NR, the baseband complexity is also a major consideration for the base stations, since an NR base station can be a small sized access node (especially at high frequencies) with a limited processing capability. At very high frequencies and large bandwidths, the receiver also has to cope with severe RF impairments.
- ***Time localization*** OFDM is very well localized in the time domain, which is important to efficiently enable (dynamic) TDD and support latency critical applications such as URLLC (see Section 2.7). Dynamic TDD is envisioned at high frequencies and provision of low latency is essential for all link types; especially backhaul and V2V links may impose very high requirements.
- ***Frequency localization*** OFDM is less localized in the frequency domain (see Section 5.5.1). Frequency localization can be relevant to support the coexistence of different services potentially enabled by mixing different waveform numerologies in the frequency domain on the same carrier (see Section 6.3.3). Frequency localization is also relevant if asynchronous access is allowed in uplink and sidelink. In general, frequency localization of a waveform may not be crucial at high frequencies where large channel bandwidths are available.
- ***Robustness to synchronization errors*** The provision of the cyclic prefix in OFDM makes it robust to timing synchronization errors (see Section 6.5.1). Robustness to the synchronization errors is relevant in scenarios where the synchronization is hard to achieve such as in the sidelink. It can also be important if asynchronous transmissions are allowed in the uplink.[3]
- ***Flexibility and scalability*** OFDM is a flexible waveform which can support diverse services in a wide range of frequencies by a proper choice of the subcarrier spacing and the cyclic prefix. (See Section 6.2 for a discussion on the OFDM numerology design that fulfills a wide range of requirements for NR.)

A summary of the waveform design requirements for different link types is shown in Fig. 6.1. A link requirement "High" for a waveform KPI tells that the given KPI is important for the given link type in general. Furthermore, a high-level assessment of OFDM is given in Table 6.1. The OFDM assessment "High" means that OFDM has good performance in general for the given KPI. We assess the D2D and V2V cases separately due to different levels of requirements. For example, V2V communication has higher requirements on the mobility and the system capacity, whereas there are lower requirements on the power efficiency when compared with UE-to-UE communication. Based on the overall assessment, it can be concluded that OFDM is an excellent choice for the NR air interface for all link types.

[3]We note that LTE only supports synchronous uplink transmission, which is realized via a timing advance mechanism at the UEs.

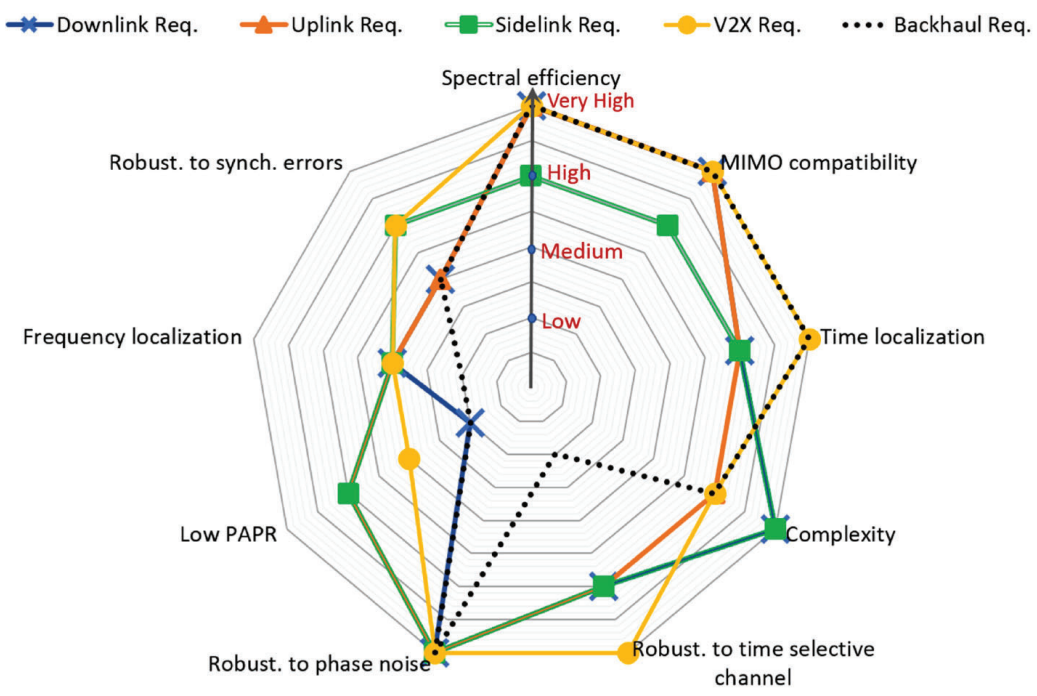


FIGURE 6.1

A high-level summary of the waveform design requirements for different link types: uplink, downlink, sidelink, V2V link, and backhaul link.

Table 6.1 A high-level assessment of OFDM

Performance indicators	OFDM assessment
Spectral efficiency	High
MIMO compatibility	High
Time localization	High
Transceiver baseband complexity	Low
Flexibility/Scalability	High
Robust. to freq.-selective chan.	High
Robust. to time-selective chan.	Medium
Medium Robust. to phase noise	Medium
Robust. to synch. errors	Medium
PAPR	High (can be improved)
Frequency localization	Low (can be improved)

6.2 SCALABLE OFDM FOR NR

CP-OFDM waveform has a few design parameters: subcarrier spacing, cyclic prefix, and the number of subcarriers. For a communication system, these parameters can be optimized based on a number of factors including carrier frequency, user mobility, phase noise, channel delay spread, quality of service

Table 6.2 Scalable OFDM numerology for 5G NR

OFDM subcarrier spacing	15 kHz	30 kHz	60 kHz	15×2^n kHz ($n \geq 3$)
OFDM symbol duration	66.67 μs	33.33 μs	16.67 μs	$66.67/2^n$ μs
Cyclic-prefix duration	4.69 μs	2.34 μs	1.17 μs	$4.69/2^n$ μs
OFDM symbol with CP	71.35 μs	35.68 μs	17.84 μs	$71.35/2^n$ μs
OFDM symbols per slot	14	14	14	14
Slot duration	1000 μs	500 μs	250 μs	$1000/2^n$ μs

Numerologies with $n = \{-2, -1\}$ are possible but not listed in this table.

requirements, signal bandwidth, and implementation complexity. For a given carrier frequency, phase noise and the Doppler effect set requirements on the minimum subcarrier spacing. The use of smaller subcarrier spacings either results in a high error vector magnitude (EVM) due to phase noise or in undesirable strict requirements on the local oscillator. Too narrow subcarrier spacings lead to performance degradations in high Doppler effect scenarios. The required cyclic-prefix overhead (and thus anticipated delay spread) sets an upper limit for the subcarrier spacing; selecting larger subcarriers would result in an undesirable high CP overhead (degrading spectral efficiency). The maximum FFT size of the OFDM modulator (or the number of total subcarriers) together with subcarrier spacing determines the channel bandwidth. Based on these relationships, the subcarrier spacing is typically chosen as small as possible while still being robust against phase noise and Doppler effect and providing the desired channel bandwidth. These design principles also apply to NR, as we will discuss in the sequel.

NR will operate from sub-1 GHz to 100 GHz using a wide range of deployment options (e.g., macro cells, micro cells, pico cells). Furthermore, NR will cover a very wide application range, including mobile broadband and several types of machine type communications. A single OFDM numerology cannot fulfill the desired frequency range and all envisioned deployment options and applications; therefore, the 3GPP has adopted a family of OFDM numerologies for NR. In the 3GPP context, OFDM numerology refers to the choice of OFDM subcarrier spacing, cyclic-prefix duration, and the number of OFDM symbols per transmission slot.[4] In particular, it has been agreed that the subcarrier spacing of OFDM can be chosen according to 15×2^n kHz, where n is an integer-valued parameter and the cyclic-prefix overhead is 7% as in LTE. The details related to NR OFDM numerologies (as agreed in the 3GPP) such as CP duration, symbol duration, and slot size, are given in Table 6.2. (We note that in 3GPP NR Release 15, the numerology is specified for up to 52.6 GHz carrier frequency with 120 kHz as the maximum subcarrier spacing as given in Table 2.1.[5]) The 15 kHz numerology is exactly the same as in LTE, with a minor exception that in LTE a slot contains seven OFDM symbols whereas a slot in NR has 14 OFDM symbols. For all numerologies, the cyclic prefix of the first OFDM symbol in every 0.5 ms interval is 16 T_s ($T_s = 32.6$ ns is the chip duration corresponding to an LTE sampling rate of 30.72 MHz) longer than the cyclic prefix of the remaining OFDM symbols in the interval. This implies that a slot length can slightly vary, depending on where it starts.

[4] See Section 2.3 for definition of the slot in NR.

[5] In 3GPP NR Release 15, the OFDM numerology is not specified for the frequency range 6 GHz to 24 GHz, since no spectrum has been identified for NR in this range. When the spectrum becomes available in the future, the numerology can be specified.

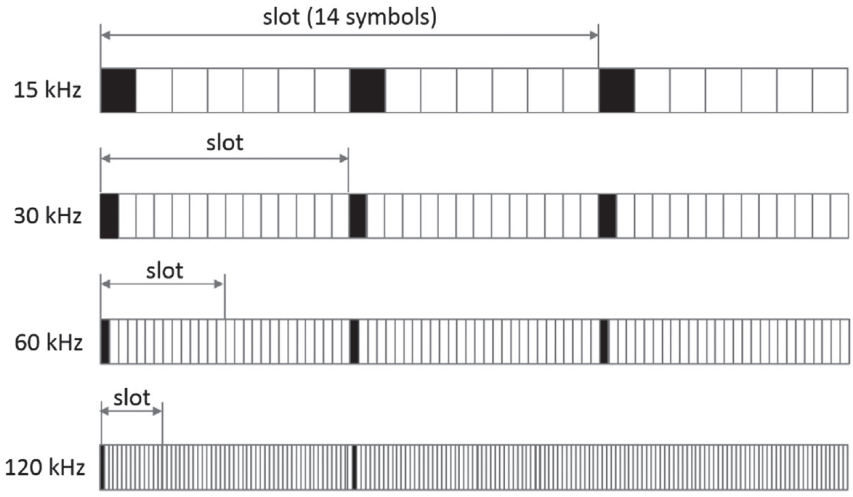


FIGURE 6.2

The NR numerology ensures symbol-wise and slot-wise time alignment. The time alignment is important for efficiently enabling TDD networks.

Fig. 6.2 illustrates OFDM symbols and slots for different numerologies (15 kHz, 30 kHz, 60 kHz, and 120 kHz), where symbols with darker colors represent OFDM symbols with longer CPs (every 0.5 ms). We observe that an integer number of slots of one numerology (with wider subcarrier spacing) fits into a slot of another numerology (with narrower subcarrier spacing). This time alignment of the slots is important for TDD networks to allow for time-aligned uplink and downlink transmission periods, which is discussed further in Section 6.2.2.

The 3GPP has introduced the concept of mini-slots in NR to support transmissions shorter than regular slot duration (see Section 2.3). A mini-slot can start at any OFDM symbol in a regular slot. Mini-slots can be useful in various scenarios—three important utilizations are shown in Fig. 6.3: i) low-latency transmissions, where the regular slot duration is too long and/or a transmission need to start immediately without waiting for the start of a slot boundary, ii) transmissions in an unlicensed spectrum, where it is beneficial to start transmission immediately after LBT, and iii) transmission in the millimeter-wave band, where the large channel bandwidths imply that the payload supported by one or two OFDM symbols can be sufficient for many of the packets. Furthermore, in the millimeter wave band the multiplexing of different users should primarily be done in the time domain to support analog/hybrid beam-forming, calling for a transmission duration shorted than a slot.

In principle, a mini-slot for a numerology can be as short as one OFDM symbol.[6] Some desirable characteristics of the mini-slot are: a mini-slot should be aligned with OFDM symbol boundaries in the regular slot; a mini-slot should end at the slot boundaries at latest; it should be possible to aggregate a mini-slot with a subsequent slot; and it may either contain control information at its beginning or at its end.

[6]In 3GPP NR Release 15, the downlink mini-slots are restricted to 2, 4, and 7 OFDM symbols.

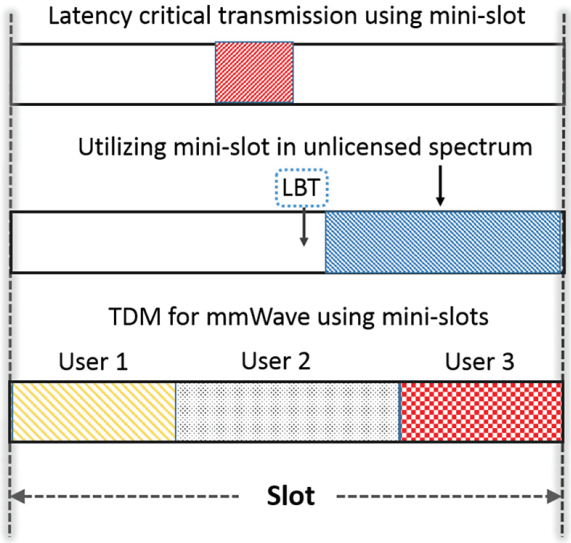


FIGURE 6.3

Mini-slots are useful in various situations, e.g., to achieve low latency, to efficiently transmit in an unlicensed spectrum based on listen-before-talk, and to schedule users at shorter time scales in the millimeter-wave band where large channel bandwidths are available.

6.2.1 WHY 15 KHZ AS BASELINE NUMEROLOGY?

The LTE OFDM numerology is 15 kHz subcarrier spacing with 7% cyclic-prefix overhead (4.69 μs). The numerology for LTE was specified after a thorough investigation in the 3GPP. For NR, it was straightforward for the 3GPP to aim for a similar OFDM numerology at LTE-like frequencies and deployments. Different subcarrier spacing options close to 15 kHz were therefore considered by the 3GPP as a baseline numerology for NR. The conclusion is to keep the LTE numerology as the baseline numerology for NR. There are two important reasons for keeping the LTE numerology as the base numerology:

- Narrow Band-IoT (NB-IoT) is a new radio access technology (already deployed since 2017) to support massive machine-type communications. NB-IoT uses the LTE numerology and provides different deployments, among others, an in-band deployment within an LTE carrier that is enabled by the selected LTE numerology. NB-IoT devices are designed to operate for 10 years or more on a single battery charge. Once such an NB-IoT device is deployed it is likely that within the device life time the embedding carrier (assuming in-band LTE deployment) gets refarmed to NR. The main reason for selecting the LTE-based numerology for NB-IoT was the option of the in-band deployment; in-band NB-IoT deployments after refarming LTE to NR would benefit from the LTE-based numerology.
- NR deployments can happen in the same band as LTE. With an adjacent LTE TDD carrier, NR must adopt the same uplink/downlink switching pattern as the LTE TDD does. Every numerology where (an integer multiple of) a subframe is 1 ms can be aligned with regular subframes in LTE. In LTE, the duplex switching happens in special subframes. To match the transmission direction in the special subframes, the same numerology as in LTE is needed.

The above arguments together with LTE numerology proven in the field were strong enough to set the baseline OFDM numerology for NR to the LTE numerology (to be used in LTE-like frequencies and deployments). This implies the same subcarrier spacing (15 kHz), the same OFDM symbol duration (66.67 μs), and the same cyclic prefix (4.69 μs).

6.2.2 WHY 15×2^n KHZ SCALING?

As discussed earlier, a set of OFDM numerologies had to be defined for NR to handle a wide range of frequencies and deployment options. These OFDM numerologies could either be unrelated to each other, i.e., the OFDM numerology for a given frequency and deployment is only based on this frequency and deployment, not considering numerologies for other frequencies and deployments at all. Another option was to define a family of OFDM numerologies which are related to each other via scaling, i.e.,

$$\Delta f_i = n_i \Delta f_{i-1}, \quad T_{cp,(i)} = \frac{T_{cp,(i-1)}}{n_i}, \tag{6.1}$$

where Δf_i and $T_{cp,(i)}$ denote the subcarrier spacing and the cyclic-prefix duration of the ith numerology and $n_i \in \mathbb{N}$ is a scaling factor. The duration of the OFDM symbol is the inverse of the subcarrier spacing. With this scaling approach, the sampling clock rates of different OFDM numerologies relate to each other via the scaling factors $\{n_i\}$, which simplifies the implementation. The 3GPP has adopted this scaling approach, i.e., the NR OFDM numerologies are derived from a baseline OFDM numerology (which is same as LTE numerology) via the scaling. In principle, the scaling factors $\{n_i\}$ could be selected independently of each other, however, it is desirable that the scaling factors follow a certain relationship to allow an efficient TDD based operation, which we discuss next.

For NR, the number of OFDM symbols per slot is equal for all numerologies, at-least up to the 120 kHz subcarrier spacing specified in 3GPP NR Release 15. Maintaining an equal number of OFDM symbols per slot for different numerologies simplifies scheduling and reference signal design. With the number of OFDM symbols per slot equal for all numerologies, the slot duration shrinks with the increase in the subcarrier spacing. This enables shorter latencies for wider subcarrier numerologies (to be used in high frequency small cell deployments where some of the low-latency applications are envisioned). Considering the equal numbers of OFDM symbols per slot for all numerologies, the following relationship holds for the respective slot durations between different numerologies:

$$T_{slot,(i)} = \frac{T_{slot,(i-1)}}{n_i} = \frac{T_{slot,(i-2)}}{n_i n_{i-1}} = \cdots = \frac{T_{slot,(1)}}{\prod_{k=1}^{i} n_k},$$

where $T_{slot,(i)}$ denotes the slot duration of the ith numerology. For adjacent TDD networks that are using different OFDM numerologies, it is desirable that an integer number of slots from one OFDM numerology fits into a slot of the other OFDM numerology to enable time-aligned downlink and uplink transmission periods. If the slot durations of different numerologies do not fulfill the above condition, then two neighboring TDD networks would require a guard time in the frame structure to enable synchronous operation, which will not be an efficient resource utilization. Therefore, the scaling factors are chosen such that a subcarrier spacing is integer divisible by all smaller subcarrier spacings, i.e.,

$$\Delta f_i = 2^{L_{(i)}} \Delta f_1, \qquad \forall i \in \{1, 2, ..., M\}, \tag{6.2}$$

where $L_{(i)} \in \mathbb{Z}$, M is the number of OFDM numerologies, and Δf_1 is subcarrier spacing of the base numerology. With this in mind, the 3GPP has specified the scaling factor as $n_i = 2^L$ in (6.1), where L is an integer.

6.3 OFDM NUMEROLOGY IMPLEMENTATION

The NR waveform is scalable—the subcarrier spacing of OFDM can be chosen according to 15×2^n kHz, where the integer-valued n is a design parameter and can be optimized for different scenarios. The 3GPP has agreed that NR should allow subcarrier spacing ranging from at least 3.75 kHz ($n = -2$) to 480 kHz ($n = 5$). A lower subcarrier spacing than 15 kHz (and a correspondingly longer cyclic prefix) is beneficial to support multicast-broadcast single-frequency network (MBSFN) transmission. For example, $n = -2$ results in a subcarrier spacing of 3.75 kHz, which is in the same range as the subcarrier spacings used in various digital broadcast standards, such as DVB, as well as being in line with some modes of NB-IoT. Hence, $n = -2$ is an option in the set of scaling factors to be considered. Choosing the value of n is not straightforward. It depends on various factors including type of deployments, service requirements, hardware impairments, mobility, performance, and implementation complexity. In the following, we provide a comprehensive discussion on these factors involved in selecting n for the NR waveform.

6.3.1 PHASE NOISE

Typically, phase noise increases with frequency of the local oscillator. It can be a major hardware impairment for NR deployments in high carrier frequencies, for example in the millimeter-wave band. Multicarrier waveforms are in general sensitive to phase noise. In an OFDM system, phase noise produces two types of degradations: a common phase error (CPE) and an intercarrier interference (ICI). We will discuss these in the following.

Consider a received baseband signal subject to phase noise $r[n] = x[n]e^{j\phi[n]}$, where $\phi[n]$ is a random process. Typically, in Phase-Locked-Loop (PLL)-based oscillators, $\phi[n]$ is a Wide-Sense Stationary (WSS) random process and $|\phi[n]| \ll 1$ [11], which leads to the following approximation:

$$r[n] = x[n]e^{j\phi[n]} \approx x[n]\,(1 + j\phi\,[n])\,. \tag{6.3}$$

Under this approximation, the demodulation of subcarrier $l \in \{1, 2, \ldots, N\}$ in an OFDM system with total N subcarriers is given by

$$\begin{aligned} R_l &= \frac{1}{N}\sum_{n=0}^{N-1} r[n]\mathrm{e}^{-j2\pi\frac{l}{N}n} \\ &\overset{(a)}{=} \frac{1}{N}\sum_{n=0}^{N-1}\sum_{i=0}^{N-1} X_i \mathrm{e}^{j2\pi\frac{i}{N}n}\mathrm{e}^{-j2\pi\frac{l}{N}n}(1 + j\phi[n]) \\ &= X_l + j\frac{X_l}{N}\sum_{n=0}^{N-1}\phi[n] + j\sum_{n=0}^{N-1}\sum_{i=0,i\neq l}^{N-1} X_i \mathrm{e}^{j2\pi\frac{i-l}{N}n}\phi[n] \end{aligned}$$

$$\overset{(b)}{=} X_l\,(1+\varphi) + N_l^{ICI}, \tag{6.4}$$

where (a) follows from $x[n] = \sum_{i=0}^{N-1} X_i e^{j2\pi \frac{i}{N} n}$ and (6.3) and (b) follows by defining $\varphi := j\sum_{n=0}^{N-1} \phi[n]$ and $N_l^{ICI} := jX_l \sum_{n=0}^{N-1} \sum_{i=1, i\neq l}^{N} X_i e^{j2\pi \frac{i-l}{N} n} \phi[n]$. According to (6.4), the transmitted symbol X_l is multiplied by the term $(1+\varphi)$. This is known as the common phase error (CPE) as it causes identical phase rotations in all subcarriers. The additive noise term N_l^{ICI} refers to the intercarrier interference, which is different for different subcarriers.

The CPE is an identical phase rotation in all subcarriers within an OFDM symbol, therefore it can easily be compensated for using pilot subcarriers (reference signals). (See Section 2.5 for an introduction to different reference signals specified for NR.) For a fast varying channel, the CPE can be compensated as part of a channel estimation process. However, for a slowly varying channel, the CPE needs to be tracked and compensated for more frequently. The 3GPP has introduced the PT-RS in NR mainly for the CPE compensation. In addition, the DM-RS whenever present, can also be used for the CPE compensation. Both the DM-RS and the PT-RS are discussed in Section 2.5.

The ICI is an additive noise (not always Gaussian) and often hard to compensate for, depending on how fast the phase noise variations are. Although the PT-RS can also be utilized for the ICI compensation, it typically requires denser pilot allocation which can degrade the system throughput.[7] An alternative is to chose the waveform numerology (the subcarrier spacing) such that there is sufficient robustness to the ICI. For a typical oscillator, the impact of the ICI decreases as the subcarrier spacing of the signal increases, as discussed in Section 5.5.5. Therefore, an OFDM system can be made robust against the ICI by choosing a larger subcarrier spacing.[8] A too low choice of the subcarrier spacing either puts strict requirements on the local oscillator or phase noise puts an upper limit on the achievable performance under the ICI. In Section 5.5.5, we have quantified the signal-to-interference ratio (SIR) due to the ICI in an OFDM system (cf. (5.24)). Assuming a perfect CPE compensation, the achievable signal-to-interference ratio (SIR) over the demodulated subcarrier l in an OFDM symbol (subject to phase noise) is given by

$$\mathrm{SIR}_l = \frac{1}{\left(\int_{-0.5}^{+0.5} S_\phi(\nu)\, W_l(\nu)\, d\nu\right)}, \tag{6.5}$$

where

$$W_l(\nu) := \sum_{i \in I_{ICI}} \frac{1}{N^2} \left| \frac{\sin\left(\pi\left(\nu - \frac{i-l}{N}\right) N\right)}{\sin\left(\pi\left(\nu - \frac{i-l}{N}\right)\right)} \right|^2. \tag{6.6}$$

In Fig. 6.4 we show the achievable SIRs for different subcarrier spacings based on a PLL based phase-noise model presented in Section 4.2.3. This phase-noise model was developed in mmMAGIC

[7] The density of the PT-RS is configurable in both time and frequency domains.

[8] It is noteworthy that the value of the CPE is constant only within one OFDM symbol. The value of CPE varies from one OFDM symbol to another OFDM symbol and its variance can increase with the subcarrier spacing (in contrast to the ICI). Therefore, an OFDM system becomes more robust against phase noise by employing the larger subcarrier spacings only if the CPE compensation is in place, which is the case for NR.

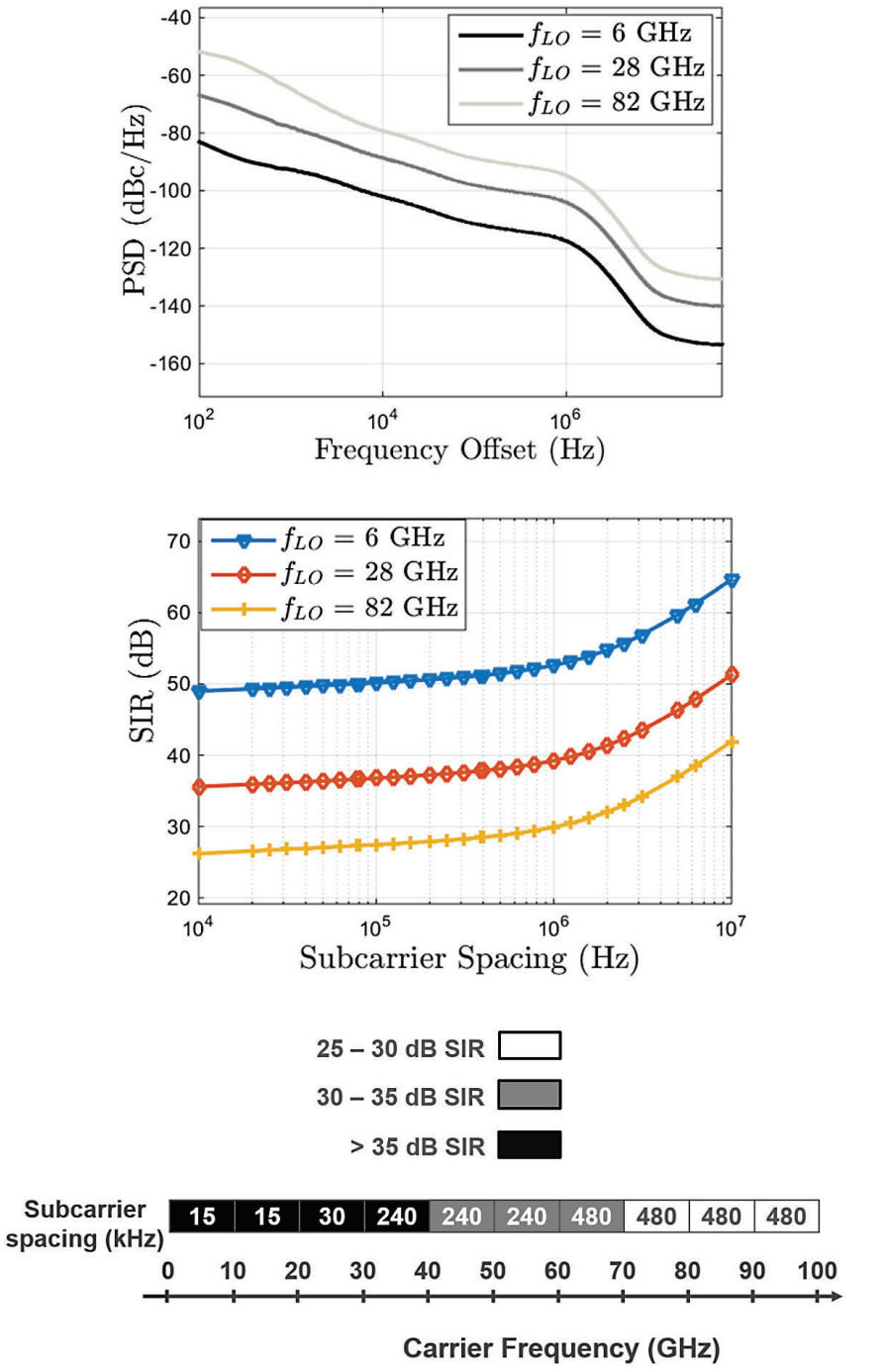


FIGURE 6.4

Achievable Signal-to-Interference Ratio (SIR) in a 100 MHz OFDM system subject to phase-noise for different subcarrier spacings. The CPE is assumed to be perfectly compensated for. (The phase-noise model is developed within mmMAGIC project.)

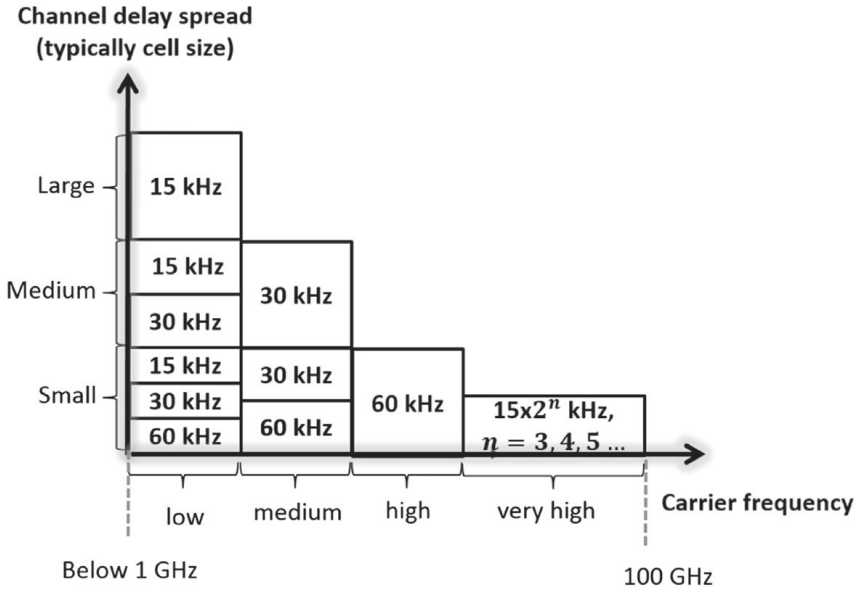


FIGURE 6.5

The NR numerology for wide range of frequencies and deployment types.

project[9] [7]. The phase-noise power spectral densities and the corresponding achievable SIRs are shown for three oscillator frequencies (6 GHz, 28 GHz, and 82 GHz). Towards the bottom of Fig. 6.4, we further specify which subcarrier spacings achieve the SIR values above 25 dB for carrier frequencies up to 100 GHz (assuming the mmMAGIC phase-noise model).

6.3.2 CELL SIZE, SERVICE LATENCY, AND MOBILITY

In an OFDM system, the cyclic prefix is chosen larger than the channel delay spread to avoid inter-symbol interference and complex equalization. This means that for a given cyclic-prefix overhead, the channel delay spread sets a lower limit on the subcarrier spacing. Typically, the delay spread reduces with cell size and so does the required cyclic-prefix duration, meaning that wider subcarrier spacings (having shorter cyclic prefix) are more suitable for deployments with smaller cell size. This goes hand-in-hand with the fact that smaller cell sizes are envisioned at higher carrier frequencies due to harsh propagation characteristics and wider subcarrier spacing that make the system robust to phase noise (which increases with oscillator frequency). Numerologies with wider subcarrier spacings are also suitable for supporting low-latency services, since the transmission slot duration (as defined in Table 6.2) is inversely proportional to the subcarrier spacing. This implies that low latencies can be achieved in small cells using a numerology with wider subcarrier spacing. Fig. 6.5 and Fig. 6.6 illustrate the relationships between cell size, carrier frequency, and achievable latency for NR.

[9] mmMAGIC was an EU project (funded by the H2020 program) that developed radio interface concepts and solutions for above 6 GHz mobile radio communications. https://5g-mmmagic.eu/.

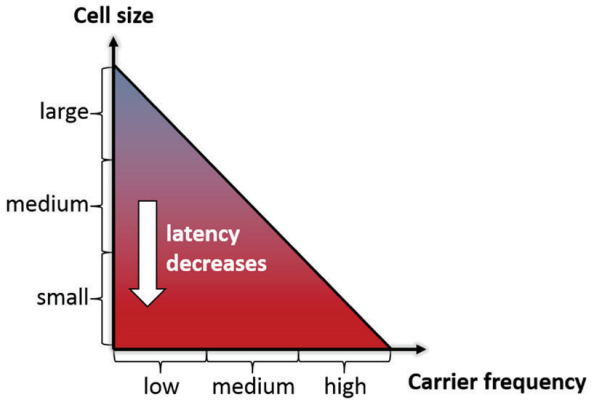


FIGURE 6.6

The relationship of achievable latency, carrier frequency, and deployment type for NR.

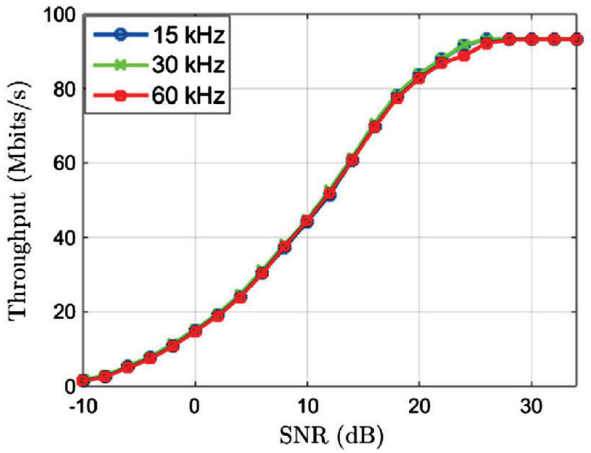


FIGURE 6.7

TDL-A channel model with 300 ns RMS delay spread.

Low latency can also be achieved in larger cells using wider subcarrier spacing, however, this can cost loss in performance either in terms of reliability or throughput. To understand this, we evaluate achievable throughput for three numerologies (15 kHz, 30 kHz, and 60 kHz) assuming the 7% CP overhead (as specified in 3GPP) subject to the Tapped Delay Line-A (TDL-A) channel model [14] with two different delay spread values (300 ns and 1000 ns). The results are shown in Fig. 6.7 and Fig. 6.8, respectively. In these simulations, we have assumed 1 ms TTI (i.e., 14 OFDM symbols for 15 kHz numerology, 28 OFDM symbols for 30 kHz numerology, and 56 OFDM symbols for 60 kHz numerology). Furthermore, we have considered link adaptation, with 64 QAM as the highest modulation order. In Fig. 6.7, we observe that all numerologies perform similarly, meaning that one can reduce the CP duration to four times smaller than what we have in LTE or one can also enable wider subcarriers to support latency critical services. Fig. 6.8 shows that the 60 kHz numerology should not

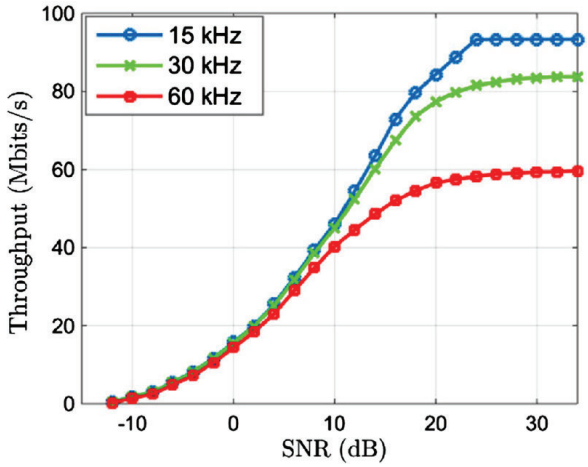


FIGURE 6.8

TDL-A channel model with 1000 ns RMS delay spread.

be used in delay spread intensive environments. One can use 60 kHz subcarrier spacing in the delay spread intensive environments with an extended cyclic prefix to support very low latency machine-type communications. However, extending the CP duration means higher overhead, and hence reduced throughput. The 3GPP has included an option of extended CP with 25% overhead for 60 kHz subcarrier spacing in NR Release 15. It is, however, important to note that the mini-slot based transmission is an alternative method to achieve very low latency without increasing the CP overhead (as discussed in Section 6.2).

In an OFDM system, the relative mobility between a user and an access node causes intercarrier interference (channel time variations or the Doppler effect leads to loss of subcarrier orthogonality). On the one hand, the Doppler effect increases with increasing carrier frequency for a given mobile scenario. On the other hand, the cell size is expected to reduce at higher carrier frequencies for 5G deployments, which implies UEs with low mobility (speeds) relative to the base station (access node). If small cells are realized at high carrier frequencies, mobility may not be a limiting factor in choosing numerology. However, for high-speed scenarios (e.g., V2V) in the millimeter-wave band, the Doppler effect can set a lower limit on the choice of the subcarrier spacings.

6.3.3 MULTIPLEXING SERVICES

Multiple services with different requirements (eMBB, mMTC, URLLC) can be efficiently supported on the same carrier by either using mini-slots or by multiplexing different OFDM numerologies in the frequency domain. For example, the mini-slot can be used to support a URLLC service within the regular slot that supports an eMBB service. Alternatively, a wideband subcarrier numerology and a narrowband subcarrier numerology can be multiplexed to support URLLC and MBB services, respectively. Fig. 6.9 illustrates these two options. Mini-slot is the preferred option due to the following drawbacks of the mixed numerology.

In an OFDM system with two (or more) different numerologies (subcarrier bandwidth and/or cyclic-prefix length) multiplexed in the frequency domain, only subcarriers within a numerology are

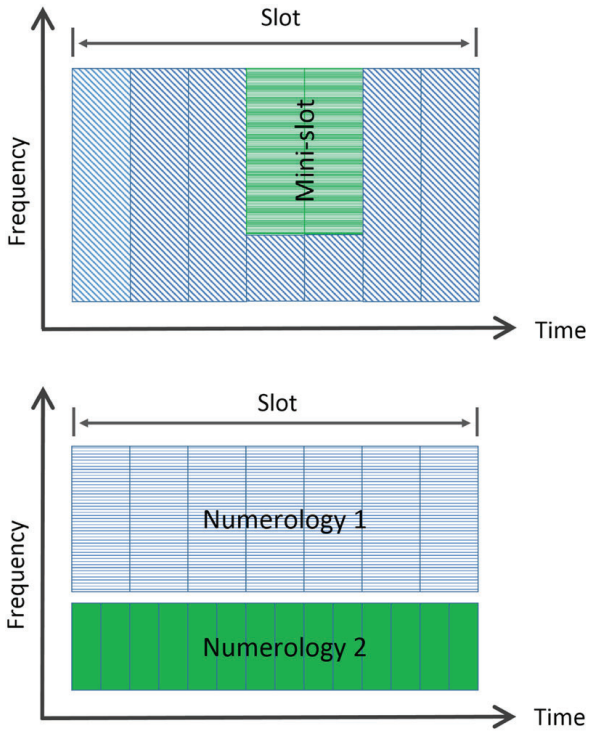


FIGURE 6.9

The mini-slot and the mixed numerology are two alternative tools for multiplexing services with different requirements on the same carrier.

mutually orthogonal. The subcarriers from one numerology interfere with the subcarriers of the other numerology, since energy leaks outside the subcarrier bandwidth and is picked up by subcarrier filters of the other numerology, as illustrated in Fig. 6.10. To reduce the internumerology interference, the transmit spectrum of each numerology must be better confined and a guard band is typically required. Moreover, an extended CP may be required with wider subcarrier spacings, depending on the channel delay spread.

6.3.4 SPECTRAL CONFINEMENT

Spectral confinement is important to reduce out-of-band (OOB) emissions. The 3GPP specifies the OOB emission requirements[10] for both base stations and devices. The better the spectrum roll-off of the signal, the easier it is to fulfill these requirements. The spectrum of the OFDM signal decays rather slowly (see Chapter 5) and the specified OOB emission requirements cannot be fulfilled without inserting guardband[11] and performing additional filtering/windowing operations. The spectrum roll-off of the OFDM signal becomes steeper as its subcarrier spacing reduces. As an example, PSDs of two

[10]The OOB emission requirements are specified in terms of spectrum emission mask and adjacent channel leakage ratio.
[11]LTE occupies 90% of the channel bandwidth. In 3GPP Release 15, the spectrum utilization for NR is above 94%.

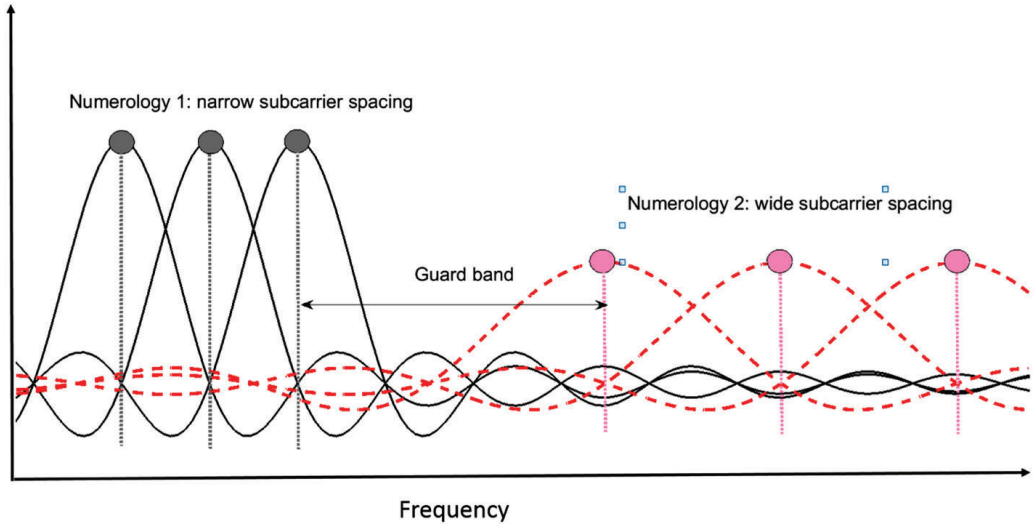


FIGURE 6.10

An illustration of the internumerology interference.

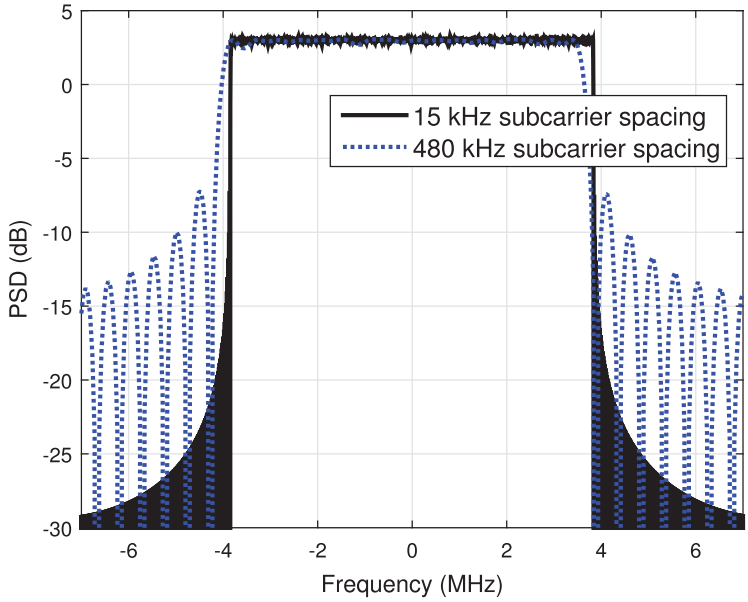


FIGURE 6.11

Comparison of OOB emissions of OFDM signals with 15 kHz and 480 kHz subcarrier spacing.

equal bandwidth OFDM signals with subcarrier spacing of 15 kHz and 480 kHz are shown in Fig. 6.11. This speaks in favor of avoiding too large subcarrier spacings in NR.

Spectral confinement is also important for reducing the internumerology interference. This can be understood with the help of Fig. 6.12 that shows two subbands with different numerologies multiplexed

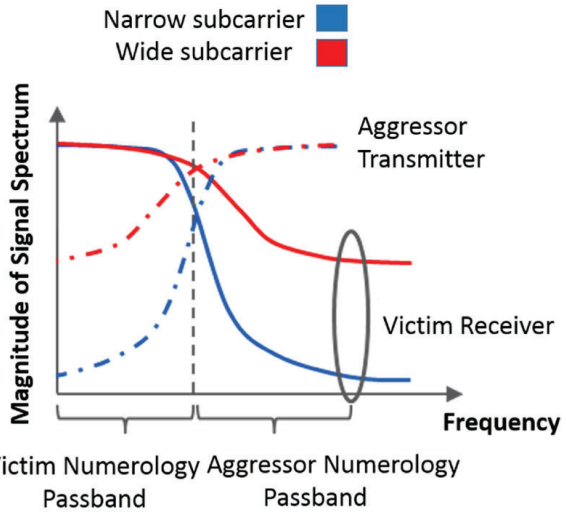


FIGURE 6.12

Both the transmitter and the receiver filtering functionality must be improved to reduce the internumerology interference efficiently.

in the frequency domain. The aggressor numerology (dash-dotted lines) must apply a spectrum emission confinement technique to reduce the energy transmitted in the passband of the victim numerology (blue[12] dash-dotted line). However, the emission control alone is not sufficient, since a victim receiver without steeper roll-off (solid red[13] curve) picks up high interference from the passband of the aggressor numerology. Only if the victim receiver (solid blue curve) and the aggressor transmitter (dashed blue curve) have improved filter functions, the internumerology interference is efficiently reduced.

Windowing and filtering are well-known techniques to confine spectrum of the OFDM signal [1,12]. Windowing has lower implementation complexity than filtering. A low complexity windowing operation can significantly improve spectrum confinement at the expense of consuming a small portion of the CP [20]. In Fig. 6.13, we compare throughput performance of OFDM without mixed numerology and W-OFDM[14] in a mixed numerology scenario where a 15 kHz numerology is multiplexed with a 60 kHz numerology. We assume that the 15 kHz numerology has 6 PRBs[15] and the 60 kHz numerology has 21 PRBs. Fig. 6.13 shows the throughput achieved by the 15 KHz numerology (i.e., considering the 60 kHz numerology as the aggressor and the 15 kHz numerology as the victim). We have used the TDL-A channel model with 1000 ns RMS delay spread (which is larger than what is typically experienced in most scenarios) and up to 64-QAM modulation for link adaptation. The two numerology subbands are separated by a guard band of 120 kHz (i.e., eight 15 kHz subcarriers). This result demonstrates that the throughput performance of W-OFDM with interference from the aggressor is close to CP-OFDM without any interference from the aggressor, even under extreme conditions with very large channel delay

[12] Dark gray in print version.
[13] Light gray in print version.
[14] W-OFDM is introduced in Section 5.1.2.2.
[15] A physical resource block contains 12 subcarriers in NR, like LTE.

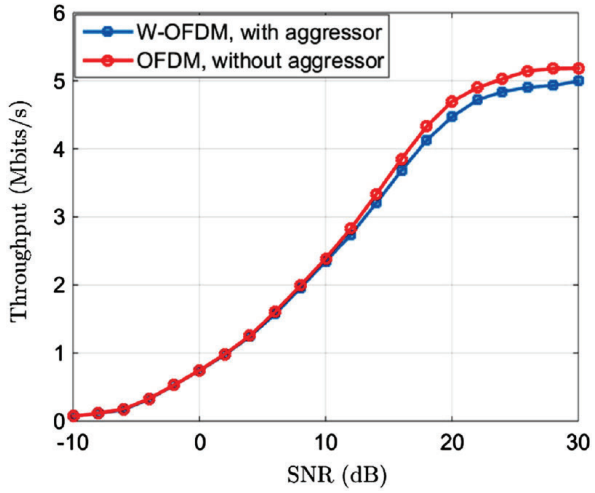


FIGURE 6.13

Comparison of W-OFDM with interference from the aggressor numerology and OFDM without interference from the aggressor numerology on TDL-A channel with 1000 ns RMS delay spread. The victim numerology is narrowband (15 kHz subcarrier spacing with 6 PRBs), the aggressor numerology is wideband (60 kHz subcarrier with 21 PRBs), and the guard band is 120 kHz.

spread, narrowband victim numerology, and wideband aggressor numerology. Hence, windowing on top of CP-OFDM is a low complexity viable technique to multiplex multiple numerologies in NR.

6.3.5 GUARD BAND CONSIDERATIONS

Guard tones can be inserted between numerologies to reduce internumerology interference and/or relax the degree of required spectrum confinement. Adding guard tones slightly increases the overhead. In a 20 MHz system with 1200 subcarriers, one guard tone corresponds to less than 0.1% overhead. It may not be worth the effort to reduce the guard band to an absolute minimum, since it increases the requirements on the spectrum confinement technique (both at transmitter and receiver) and also complicates other system design aspects, as discussed in the following.

Consider Fig. 6.14, where one narrowband subcarrier is inserted as a guard tone between numerology 1 (15 kHz subcarrier spacing) and numerology 2 (60 kHz subcarrier spacing). A physical resource block in NR contains 12 subcarriers for all numerologies. If the scheduling is done as indicated for numerology 2, then the subcarriers of numerology 2 are not on the 60 kHz resource grid (the first subcarrier of the blue (light gray in print version) resource block is on narrow subcarrier 41, which corresponds to the wide subcarrier 10.25, i.e., there is a fractional subcarrier shift).

To avoid such fractional subcarrier shift, subcarrier frequencies in each numerology should coincide with the natural grid of the numerology $n \times \Delta f$, with Δf the subcarrier spacing of the numerology. However, even with this requirement the wide resource blocks (numerology 2) are still not on their natural grid if compared to cell 2; see Fig. 6.15. Such a misaligned resource grid implies that all users of numerology 2 would have to be dynamically informed as regards this offset (since this offset depends on the scheduling decision). In another cell, a different offset may be present or another cell

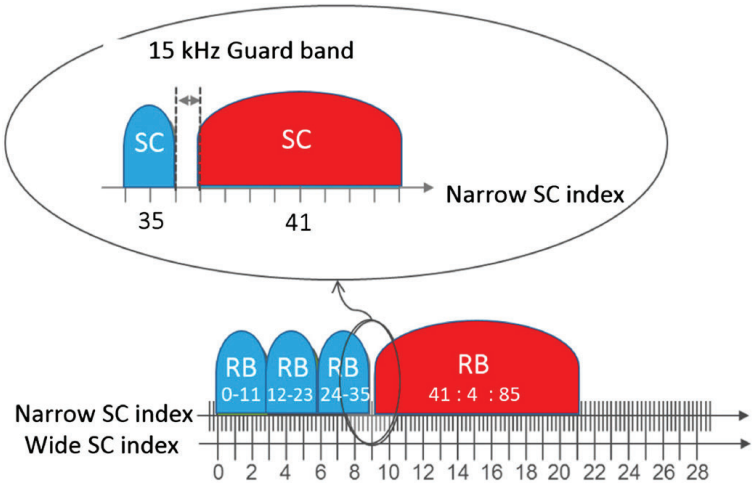


FIGURE 6.14

One narrowband subcarrier (SC) is inserted as a guard between numerology 1 (15 kHz) and numerology 2 (60 kHz). The first subcarrier of numerology 2 is located at 41 × 15 kHz, which corresponds to subcarrier 10.25 in a 60 kHz subcarrier grid.

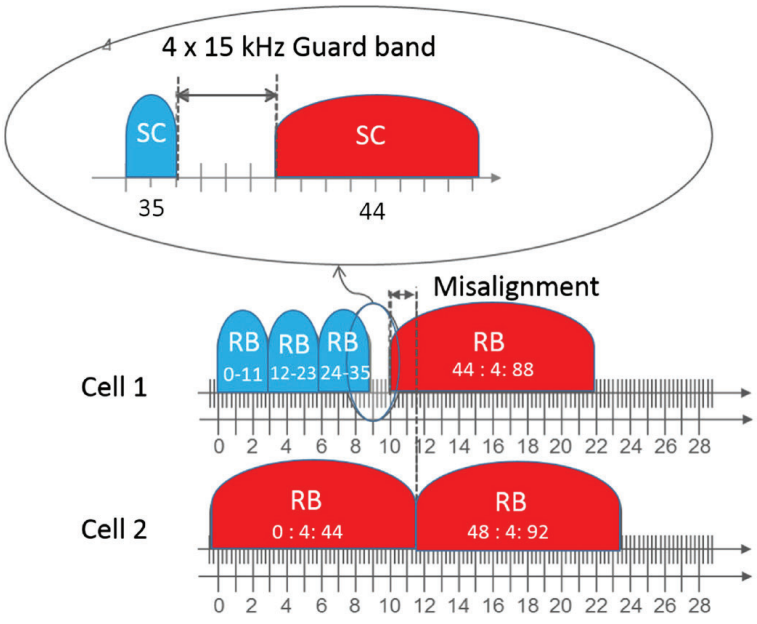


FIGURE 6.15

Four narrowband subcarriers are inserted as a guard between numerology 1 (15 kHz) and numerology 2 (60 kHz). The subcarriers of numerology 2 are located on its natural resource grid. However, numerology 2 resource blocks are still misaligned across cells.

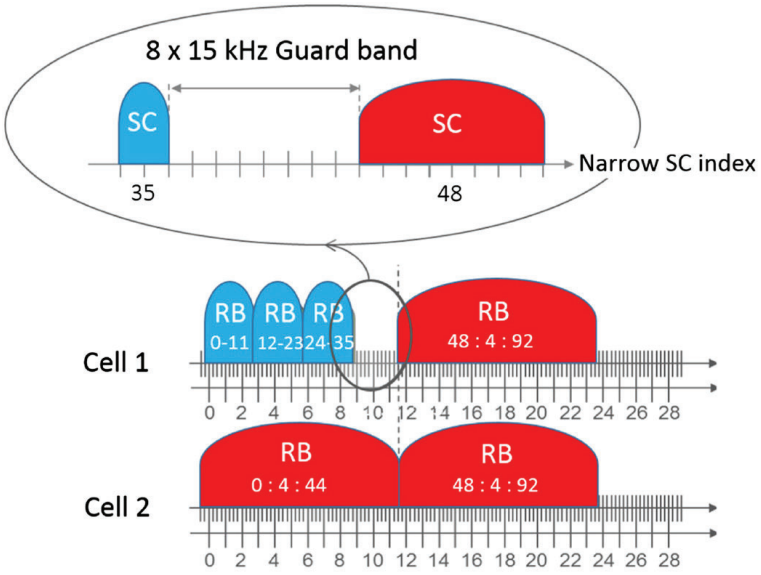


FIGURE 6.16

Eight narrowband subcarriers are inserted as guard between numerology 1 (15 kHz) and numerology 2 (60 kHz). The subcarriers of numerology 2 are located at its natural resource grid and numerology 2 resource blocks are aligned across cells.

may only operate with numerology 2 (as shown in Fig. 6.15). The resource blocks in different cells would not be aligned making intercell interference coordination (ICIC),[16] creation of orthogonal reference signals across cells, and interference prediction across cells more difficult. Alternatively, the first red (dark gray in print version) resource block in cell 1 in Fig. 6.15 could be a fractional resource block (corresponding to the bandwidth marked by misalignment). Special definitions of the reference signals and rate matching would be required for all possible fractional resource blocks. For the fractional resource block in cell 1 and the overlapping resource block in cell 2, the same disadvantages exist.

The cleanest solution is to limit the location of the resource blocks to their natural resource grids. For example, as shown in Fig. 6.16, numerology 1 (15 kHz) resource blocks always start at a frequency of $n \times 12 \times 15$ kHz and numerology 2 (60 kHz) resource blocks start at a frequency of $n \times 12 \times 60$ kHz. This simplifies the ICIC, makes the interference predication across cells easier, and enables orthogonal reference signals of the same numerology across cells. For the (15 kHz, 60 kHz) numerology combination, the resulting guard band is eight narrowband (15 kHz) subcarriers. In a 20 MHz system with around 1200 narrowband subcarriers, the spectral loss due to this guardband is less than 1%.

[16]The ICIC refers to the coordination between neighboring base stations to avoid severe interference situations. The overall system efficiency and user experience can be significantly improved with the ICIC, especially in heterogeneous network deployments with overlapping layers of base stations with large differences in the downlink transmission power.

6.3.6 IMPLEMENTATION ASPECTS

OFDM is implemented in baseband via the fast Fourier transform (FFT) operation (see Fig. 5.7). The FFT/IFFT size should be selected considering the subcarrier spacing, the channel bandwidth, and the affordable computational complexity. The asymptotic complexity of an N-point FFT operation is $\mathcal{O}(N \log N)$. For a given subcarrier spacing, the FFT size determines the maximum channel bandwidth (without carrier aggregation). The maximum channel bandwidth in LTE is 20 MHz, with 2048 FFT size, 30.72 MHz sampling rate, and 1200 is the number of active carriers (100 PRBs). In 3GPP NR Release 15, the maximum number of active subcarriers is 3300 and the largest currently defined channel bandwidth occupies 3276 subcarriers. This can be implemented using an FFT size of 4096. Note that a smaller FFT size may be used if the number of active subcarriers is less.

6.4 IMPROVING POWER EFFICIENCY OF NR WAVEFORM

A major drawback of OFDM waveform is its high PAPR. A high PAPR in the OFDM system causes the power amplifier to operate in a nonlinear region, which contributes to a spectral growth in the form of intermodulation between subcarriers and out-of-band emissions, as described in Section 4.1. Therefore, to keep the nonlinear effects within certain limits, the power back-off is commonly used in the power amplifier, which reduces the coverage. Another possibility is to increase the linear region of the power amplifier, which in turn results in larger amplifiers and hence, higher power consumption. A very common solution is to deploy DPD algorithms to cancel out the nonlinear distortion from the amplifier. This, however, is a high complexity solution, in particular for large antenna arrays.

The power efficiency of the waveform can be very important in the uplink and the sidelink, for coverage limited UEs. A high PAPR (with a high power back-off requirement) mainly affects the highest output transmission power level for the UE. Hence, a low PAPR is more relevant to the coverage-limited UEs. In LTE, the downlink uses OFDM, while the uplink and the sidelink use DFTS-OFDM due to its lower PAPR than OFDM. (See Section 5.2 for an introduction to DFTS-OFDM.) For NR, the 3GPP has specified DFTS-OFDM as an optional[17] waveform for the uplink, however, without spatial multiplexing (MIMO). A MIMO DFTS-OFDM transceiver has a significantly higher implementation complexity than OFDM and worse performance than OFDM. Furthermore, in the coverage limited scenarios spatial multiplexing is often not useful. Typically, UEs at the cell edge use only a small number of resource blocks (i.e., narrow bandwidth allocation) for transmission with low spectral efficiencies (low code rate and low order modulation formats, e.g., QPSK). Due to these reasons, NR supports DFTS-OFDM only for a single stream transmission in the uplink.

When a UE is in good cell coverage and has a wide bandwidth allocation, a high PAPR is often not critical. In this scenario, OFDM is the natural choice. Using OFDM instead of DFTS-OFDM in the uplink (and also in the sidelink) has several advantages:

- MIMO transmission is one of the key features of NR and it is expected that even uplink MIMO will be widely adopted for the enhanced mobile broadband usage (see Chapter 7). With MIMO

[17] In 3GPP NR Release 15, DFTS-OFDM is optional in the uplink and it is up to the base station to select which waveform (OFDM or DFTS-OFDM) should be used in the uplink. This implies that an NR capable UE need to implement both waveforms and a gNB (or base station) can choose to implement only OFDM.

transmission, OFDM is known to provide a significant superior link performance compared to DFTS-OFDM [8].
- DFTS-OFDM requires an additional DFT precoder in the transmitter and an extra IDFT decoder and equalizer in the receiver which increases the baseband complexity, significantly with the MIMO transmissions.
- OFDM is more flexible than DFTS-OFDM in terms of scheduling users' data, control signals, and reference signals, resulting in an enhanced system capacity.
- Having the same transmission scheme in both uplink and downlink makes the whole system design symmetric. A single waveform (i.e., OFDM) for all link types (uplink, downlink, sidelink, backhaul link) simplifies the design by avoiding the need for separate baseband implementations for different waveforms.

Although using OFDM in the uplink comes with major advantages, yet it suffers from the problem of the high PAPR which should be addressed if it is to be employed in all scenarios. There exist several PAPR reduction techniques for OFDM with their pros and cons [10]. While choosing an appropriate PAPR reduction technique, several factors need attention. For example, the PAPR reduction capability, implementation complexity, transparency between the transmitter and the receiver sides, and link performance degradation. Some of the techniques require implementation of multiple FFTs and/or transmission of side information to the receiver, which may be either undesirable or not possible.[18] In the following, we provide a brief overview of some well-known PAPR reduction schemes based on [10]. Broadly speaking, one can divide the PAPR reduction methods into two groups: i) techniques with distortion, and ii) distortion-less techniques.

6.4.1 TECHNIQUES WITH DISTORTION

The following schemes distort the OFDM signal while reducing its PAPR:

- **Clipping** The most simple and widely used technique is to limit the signal envelope to a predefined threshold by means of an amplitude clipper. The amplitude clipping function is given by

$$f\,(x[n]) = \left\{ \begin{array}{c} x[n], \quad |x[n]| \leq A_{max}, \\ A_{max}\exp(j\phi(x[n])), \quad |x[n]| > A_{max}, \end{array} \right\} \tag{6.7}$$

  where A_{max} is the amplitude threshold and $\phi(x[n])$ is the phase of $x[n]$. Clipping is a nonlinear transformation of the signal. It causes both in-band and out-of-band distortions. The out-of-band distortion can be reduced by performing filtering after clipping. However, the filtering operation may increase the signal amplitude above the clipping threshold. To address this issue, repeated clipping and filtering operations may be employed at the cost of increased implementation complexity. The in-band distortion cannot be improved by filtering and therefore degrades the BER performance and the spectral efficiency.
- **Companding** Companding refers to a nonlinear transformation of a signal in which lower signal amplitude levels are increased and peak signal amplitudes remain unchanged. The average signal

[18] In 3GPP NR Release 15, any operation performed on CP-OFDM at the transmitter side has to be receiver agnostic.

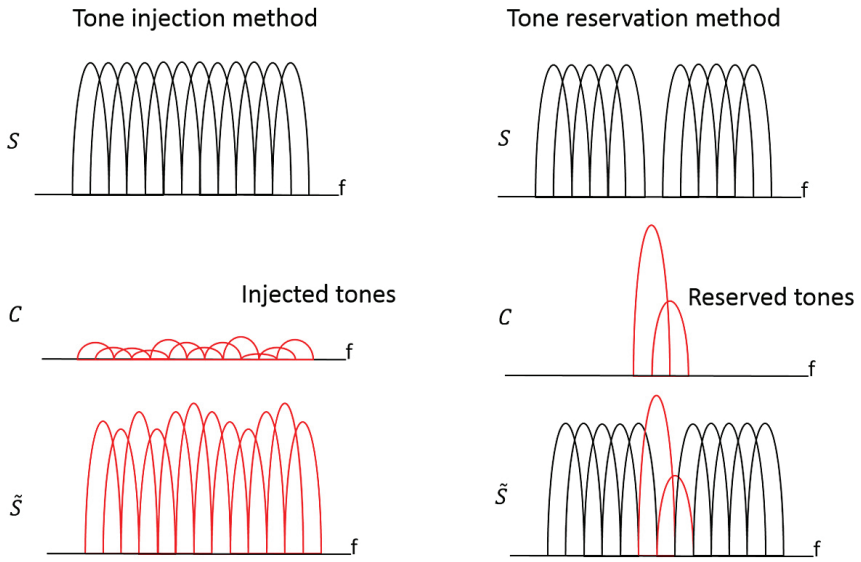


FIGURE 6.17

An illustration of Tone Injection (TI) and Tone Reservation (TR) methods.

power increases after the companding operation, which reduces PAPR. Several companding transformations exist in the literature, for example, exponential companding and polynomial companding are widely used. Companding is an invertible transformation since the companding function is a monotonically increasing function, unlike clipping. This means that a companded signal can be recovered at the receiver by applying an inverse transformation. Both the clipping and the companding methods have low implementation complexity, however, they give rise to relatively high spectral emissions.

- **Tone Injection (TI)** Tone injection refers to superimposing additional subcarriers with optimized amplitude and phases on the information bearing data subcarriers. An illustration is shown in Fig. 6.17 for the injection of the additional tones C into the data subcarriers S to get a PAPR reduced signal $\tilde{S} = S + C$. This is equivalent to expanding the constellation size of the signal, i.e., each point in the original constellation can be mapped in different points in the expanded constellation. As each data symbol can be mapped to one of the several equivalent constellation points, the transmitter can use this degree of freedom to reduce the PAPR. In this case, the receiver needs to know how to map the redundant constellations to the original constellation.
- **Active Constellation Extension (ACE)** ACE is similar to TI, but with the difference that in ACE only outer constellation points are dynamically extended away from the original constellation. Constellation extension gives additional degrees of freedom to reduce the PAPR by optimizing the constellation points. By extending the outer constellation, the spacing between constellation points also increases, which improves BER performance. A drawback of ACE is the increase in transmit power. The usefulness of ACE is restricted to the modulation schemes with large constellation size.

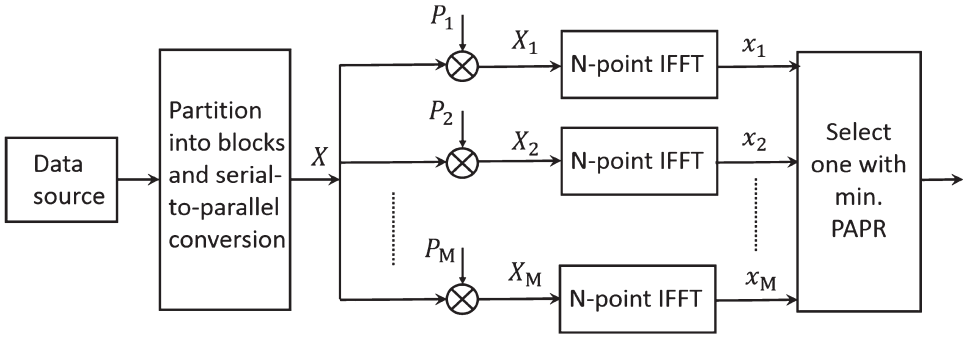


FIGURE 6.18

A schematic of Selective Mapping (SLM) method.

6.4.2 DISTORTION-LESS TECHNIQUES

The following PAPR reduction schemes do not introduce distortion:

- **Tone Reservation (TR)** In this technique, a set of subcarriers within an OFDM symbol are reserved solely for the PAPR reduction. These reserved subcarriers do not carry any data; they are modulated such that the summation of all subcarriers (the data subcarrier and the reserved subcarriers) generate a signal with a low PAPR. In Fig. 6.17, an illustration is given for the superposition of the reserved tones C with the data subcarriers S to get a PAPR reduced signal $\tilde{S} = S + C$. The procedure of finding and optimizing the reserved tones is typically of moderate complexity. The effectiveness of this method depends on the number of reserved tones.
- **Selective Mapping (SLM)** In this method, a number of data blocks carrying same information are generated and the data block that has lowest PAPR is selected for transmission, as shown in Fig. 6.18. Different data blocks are generated by multiplying modulated symbols with different phase vectors before IFFT. The information as regards phase vectors has to be transmitted to the receiver. Therefore, this scheme is not transparent to the receiver and there is a loss of spectral efficiency due to the transmission of side information. The PAPR reduction ability of this method is directly related to the implementation complexity; the larger the set of phase vectors and the number of IFFTs, the higher will be the PAPR reduction capability.
- **Partial Transmit Sequence (PTS)** In this method, the input data block is divided into a set of disjoint subblocks. Each subblock is padded with zeros, IFFT operation is performed on each sub-block, which is then multiplied with a phase vector, as shown in Fig. 6.19. The subblocks are then combined such that PAPR of the time-domain OFDM signal is minimized.
- **Suitable Coding** The idea is to select the codewords that reduce the PAPR of the transmitted signal. Peak power is achieved in the OFDM signal when N subcarriers with same phase values are super-imposed. The peak power is N times the average power. However, not all codewords result in a high PAPR. The codewords can be selected such that likelihood of occurrence of the same phase for N subcarriers can be minimized. Various coding methods have been shown effective for reducing the PAPR. A drawback of the coding-based PAPR reduction methods is the potential loss in the coding rate (i.e., lower spectral efficiency).

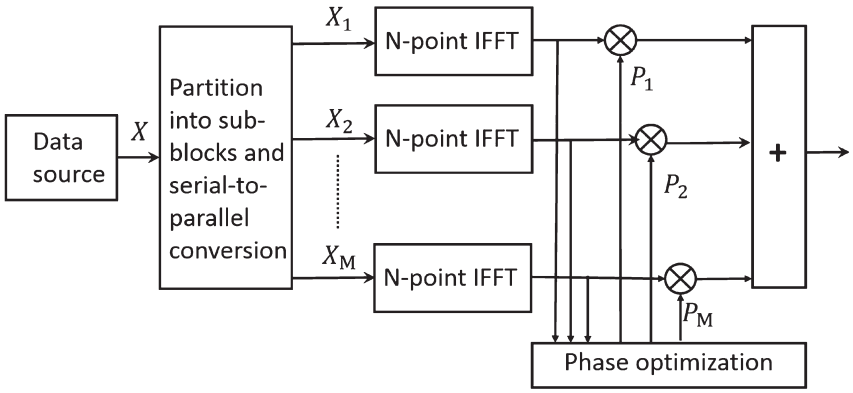


FIGURE 6.19

A schematic of Partial Transmit Sequence (PTS) method.

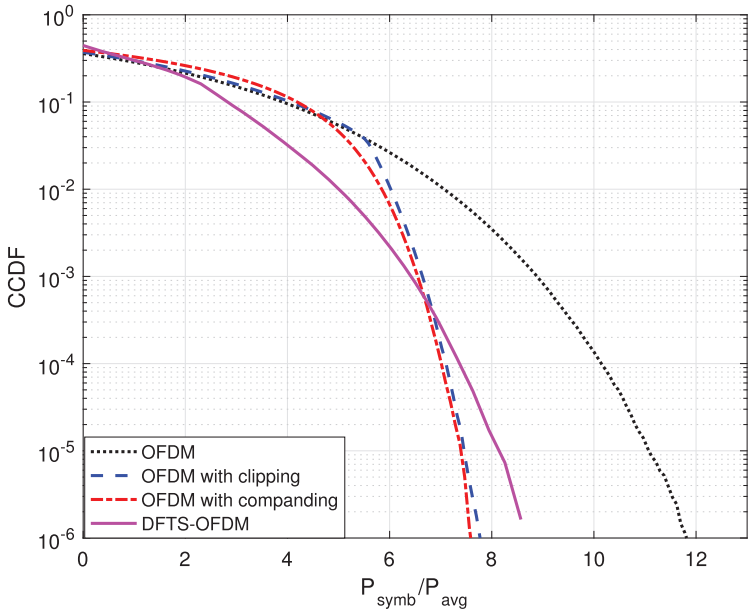


FIGURE 6.20

CCDF of PAPR (P_{symb}/P_{avg}).

Next, we look at the effectiveness of two low complexity PAPR reduction techniques—the clipping and the companding methods. We assume the following OFDM parameters: 2048 FFT size, 60 kHz subcarrier spacing, and 16 QAM modulation. The exponential companding degree is set to 1 and the clipping ratio (i.e., the ratio of the amplitude threshold to the square root of the average power of the signal) of the amplitude clipping scheme is set to 1.7, in order to have similar level of PAPR reduction with the two schemes. Fig. 6.20 shows the CCDF of the ratio of the instantaneous power of an OFDM symbol (P_{symb}) to the average power (P_{avg}) of the OFDM system with and without different PAPR

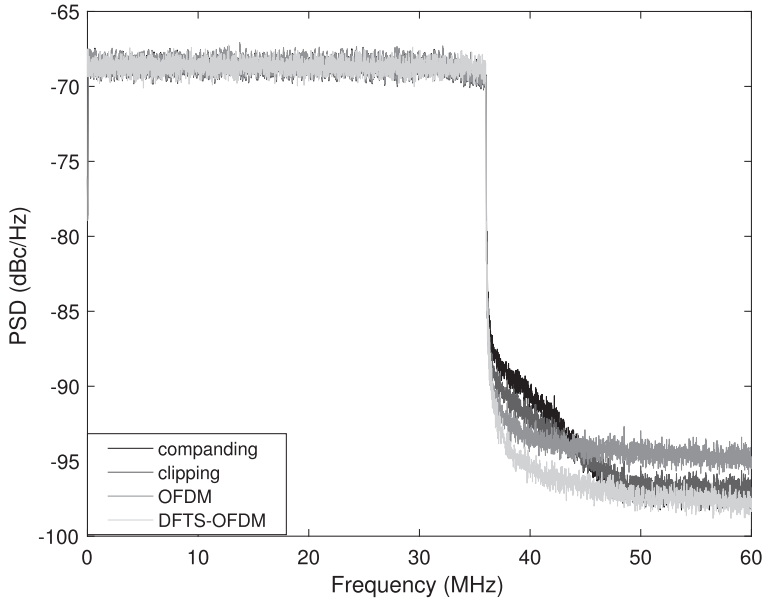


FIGURE 6.21

PSD of OFDM signals with and without different PAPR reduction schemes.

reduction schemes. For comparison, the PAPR performance of a DFTS-OFDM signal is also plotted in the same figure. The effectiveness of the clipping and exponential companding schemes in the OFDM system is clearly demonstrated. Fig. 6.21 shows the PSDs of the OFDM signals with and without the PAPR reduction schemes. For comparison, the PSD of the DFTS-OFDM signal is also included in the figure. We observe that there is some spectral regrowth when these PAPR reduction schemes are employed (which is one of the drawbacks). Nevertheless, we conclude that the high PAPR issue of OFDM can be addressed by implementation of a PAPR mitigation scheme which is not complex from the implementation point of view and achieves similar performance as DFTS-OFDM.

6.5 EFFECTS OF SYNCHRONIZATION ERRORS

In general, an OFDM system is robust to timing synchronization errors due to the presence of the cyclic prefix. However, it is sensitive to frequency synchronization errors due to its overlapping orthogonal narrowband subcarriers that can lose orthogonality when subject to frequency errors. In this section, we will briefly discuss the effects of timing and frequency synchronization errors in an OFDM system.

6.5.1 EFFECT OF TIMING OFFSET

Consider a discrete time OFDM signal with its mth OFDM symbol denoted as $x_m[n] = \sum_{i=0}^{N-1} X_{m,i} \times e^{j2\pi \frac{i}{N} n}$, having N subcarriers (generated by an N point IFFT) and N_{cp} cyclic-prefix samples. To

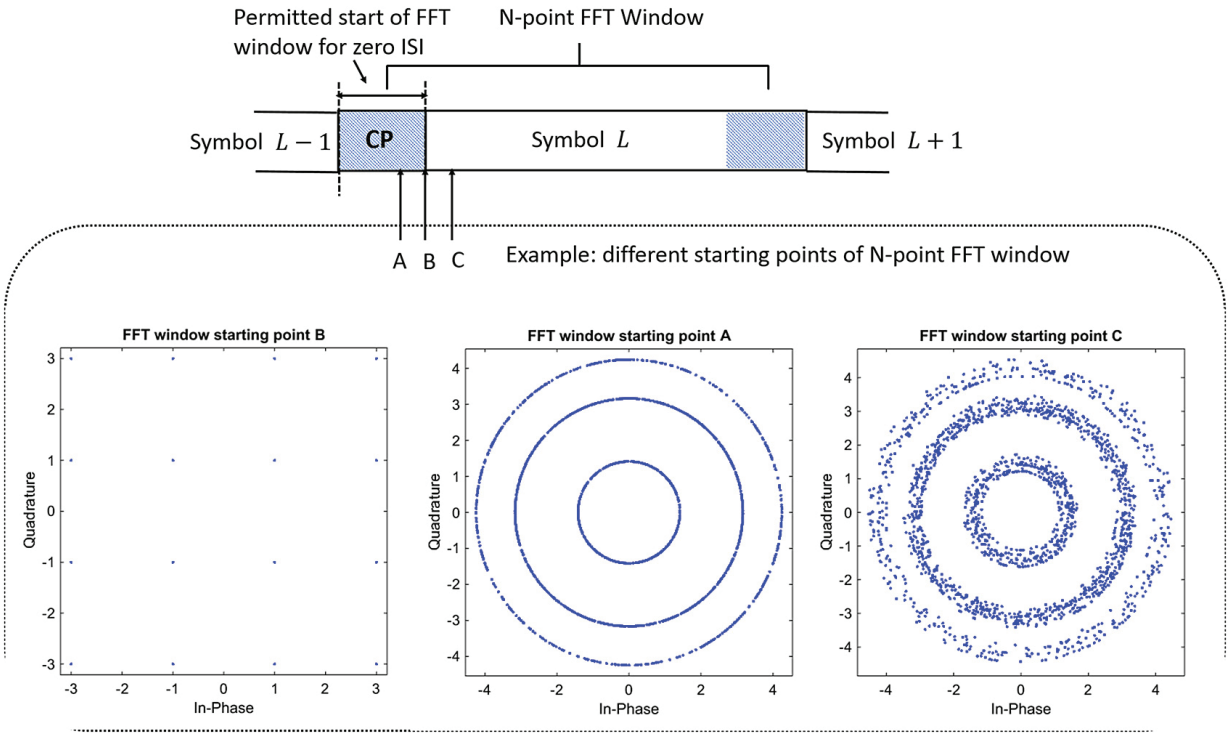


FIGURE 6.22

Effects of timing offset in an OFDM system.

demodulate this signal, an OFDM receiver must place an FFT window (perform an FFT operation) on N consecutive samples of each OFDM symbol. The ideal FFT window starting point is the first sample after the removal of the cyclic prefix, which leads to a perfect reconstruction of the transmitted modulation symbols. This is illustrated in Fig. 6.22 as the starting point B, along with the effects of starting an FFT window within the CP duration (the starting point A) and outside the CP duration, capturing portion of another symbol (the starting point C). For the starting point A (FFT window within the CP duration), the demodulated QAM symbols experience phase rotation (a frequency error). Mathematically, the demodulated subcarriers in this case are given by

$$\hat{X}_{m,k} = X_{m,k} e^{\frac{-j2\pi n_o k}{N}}, \quad k = 1, 2, \ldots, N, \tag{6.8}$$

where n_o is the timing offset. According to (6.8), the kth subcarrier experiences a phase shift $e^{\frac{-j2\pi n_o k}{N}}$. This phase shift (or the timing offset n_o) can be estimated (via pilot subcarriers) and easily compensated for in the frequency domain (post FFT). Hence, a timing offset of $1 \leq n_0 \leq N_{cp}$ is not problematic. In practice, when a signal is received over a multipath channel with maximum delay spread L, a timing offset of $1 \leq n_0 \leq N_{cp} - L$ is tolerable.

The timing offset become problematic when an FFT window captures samples of an adjacent OFDM symbol (i.e., the FFT window starts after the first OFDM sample in a CP removed symbol), illustrated by the starting point C in Fig. 6.22. In this case the demodulated subcarriers suffer from

intercarrier interference (ICI) and intersymbol interference (ISI). Mathematically, the kth demodulated subcarrier in the mth OFDM symbol subject to ICI and ISI (from $m+1$th OFDM symbol) due to the timing offset n_o is given by

$$R_{m,k} = X_{m,k}\left(1-\frac{n_o}{N}\right)\mathrm{e}^{\frac{j2\pi n_o k}{N}} + \underbrace{\frac{1}{N}\sum_{n=0}^{N-1-n_o}\sum_{\bar{k}=0;\bar{k}\neq k}^{N-1} X_{m,k}\mathrm{e}^{\frac{j2\pi(n+n_o)\bar{k}}{N}}}_{\text{ICI}}$$
$$+ \underbrace{\frac{1}{N}\sum_{n=N-n_o}^{N-1}\sum_{\bar{k}=0}^{N-1} X_{m+1,k}\mathrm{e}^{\frac{j2\pi(n-N-N_{cp}+n_o)\bar{k}}{N}}\mathrm{e}^{\frac{-j2\pi n_o k}{N}}}_{\text{ISI}}. \tag{6.9}$$

6.5.2 EFFECT OF CARRIER FREQUENCY OFFSET

The carrier-frequency offset (CFO) refers to the mismatch between the frequency of the received signal and the frequency of the local oscillator at the receiver. Two factors contribute to the CFO: i) the frequency mismatch between the transmitter and the receiver oscillators; ii) the Doppler effect due to the relative mobility of the transmitter and the receiver. In practice, the oscillators at the transmitter and the receiver can never oscillate at identical frequencies; therefore, there always exists a CFO in the received baseband signal. Due to the CFO, the baseband signal is shifted in the frequency domain. As we go higher in carrier frequencies, the CFO is more pronounced due to both the oscillator frequency mismatch and the Doppler effect.

Multicarrier waveforms are more sensitive to the CFO than single-carrier waveform, since a subcarrier bandwidth is typically much smaller than the overall bandwidth in multicarrier waveforms. A small CFO can cause significant degradation in the symbol error rate performance. In an OFDM system, the CFO produces two effects: i) a common phase error (CPE), ii) an intercarrier interference (ICI). The CPE refers to a common phase rotation in all subcarriers and the ICI refers to an interference between the subcarriers due to the loss of subcarrier orthogonality. Let us define the CPE and the ICI, mathematically. Consider a received baseband signal subject to the CFO: $r[n] = x[n]\mathrm{e}^{j2\pi\frac{\epsilon}{N}n}$, where ϵ is a normalized fractional CFO. The demodulation of the subcarrier l is then given by

$$\begin{aligned}
R_l &= \frac{1}{N}\sum_{n=0}^{N-1} r[n]\mathrm{e}^{-j2\pi\frac{l}{N}n} \\
&\overset{(a)}{=} \frac{1}{N}\sum_{n=0}^{N-1}\sum_{i=0}^{N-1} X_i\mathrm{e}^{j2\pi\frac{i}{N}n}\mathrm{e}^{-j2\pi\frac{l}{N}n}\mathrm{e}^{j2\pi\frac{\epsilon}{N}n} \\
&= X_l\frac{1}{N}\sum_{n=0}^{N-1}\mathrm{e}^{j2\pi\frac{\epsilon}{N}n} + \sum_{n=0}^{N-1}\sum_{i=0,i\neq l}^{N-1} X_i\mathrm{e}^{j2\pi\frac{i-l+\epsilon}{N}n} \\
&\overset{(b)}{=} \alpha X_l\underbrace{\mathrm{e}^{j\pi\epsilon\frac{N-1}{N}}}_{\text{CPE}} + \underbrace{\sum_{n=0}^{N-1}\sum_{i=0,i\neq l}^{N-1} X_i\mathrm{e}^{j2\pi\frac{i-l+\epsilon}{N}n}}_{\text{ICI}_l},
\end{aligned} \tag{6.10}$$

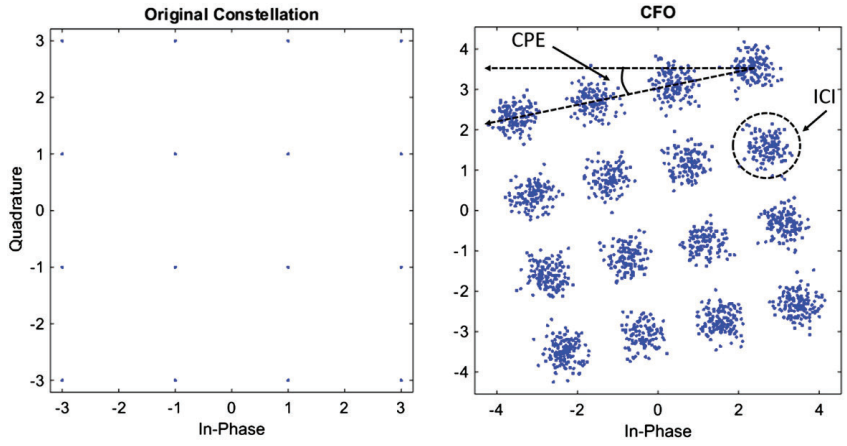


FIGURE 6.23

Effects of carrier-frequency offset in an OFDM system.

where α is an attenuation factor common to all subcarriers, which is given by

$$\alpha = \frac{sinc\,(\epsilon)}{sinc\,(\epsilon/N)}. \tag{6.11}$$

As an example, the two effects produced by CFO in an OFDM system (CPE and ICI) are illustrated in Fig. 6.23 for a 16-QAM constellation. CPE causes an identical phase rotation in all subcarriers within an OFDM symbol and ICI acts as an additive noise in the demodulated subcarrier. CPE can easily be compensated for as part of the channel equalization process. Assuming perfect CPE compensation, the achievable signal-to-interference ratio (SIR) due to ICI in the demodulated subcarrier l in the OFDM symbol subject to CFO is given by

$$\text{SIR}_l = \frac{\alpha^2 E\left[|X_l|^2\right]}{E\left[|\text{ICI}_l|^2\right]}, \tag{6.12}$$

where $E[\cdot]$ is the expectation operator.

6.5.3 SAMPLING FREQUENCY OFFSET

The sampling frequency offset (SFO) refers to a mismatch between the oscillator of the transmitter and the oscillator of the receiver. In practice, the sampling clocks of the transmitter and the receiver are offset by a few parts per million (ppms), which can cause major degradation in an OFDM system if not properly compensated for. In the following, we briefly discuss the effect of the SFO in an OFDM system.

Consider an OFDM transmitter with the sampling time T_T and an associated OFDM receiver with the sampling time T_R (i.e., assuming that the sampling rates are not identical). Then the received baseband OFDM signal can be expressed as $\sum_{i=0}^{N-1} X_i e^{j2\pi in\frac{T_R}{T_T}}$. Let us define $\eta := \frac{T_R - T_T}{T_T}$. The de-

modulation of a subcarrier l is then given by

$$\begin{aligned} R_l &= \frac{1}{N}\sum_{n=0}^{N-1} r[n]\mathrm{e}^{-j2\pi\frac{l}{N}n} \\ &= \frac{1}{N}\sum_{n=0}^{N-1}\sum_{i=0}^{N-1} X_i \mathrm{e}^{j2\pi in\frac{T_R}{T_T}}\mathrm{e}^{-j2\pi l\frac{n}{N}} \\ &= X_l\sum_{n=0}^{N-1}\mathrm{e}^{-j2\pi ln\frac{(T_R-T_T)}{T_T N}} + \sum_{n=0}^{N-1}\sum_{i=0,i\neq l}^{N-1} X_i \mathrm{e}^{j2\pi n\frac{iT_R-lT_T}{T_T N}} \\ &= \alpha_l X_l \underbrace{\sum_{n=0}^{N-1}\mathrm{e}^{j\pi l\eta\frac{N-1}{N}}}_{\text{linear phase drift}} + \underbrace{\sum_{n=0}^{N-1}\sum_{i=0,i\neq l}^{N-1} X_i \mathrm{e}^{j2\pi\frac{i-l+\epsilon}{N}n}}_{\text{ICI}}, \end{aligned} \tag{6.13}$$

where α_l is given by

$$\alpha_l = \frac{sinc\,(l\eta)}{sinc\,(l\eta/N)}. \tag{6.14}$$

From (6.13), we observe that the SFO induces a linear phase drift across the subcarriers in contrast to CFO that produces a common phase rotation (CPE) in all subcarriers. Moreover, different subcarriers are also attenuated differently when subject to the SFO, unlike the CFO. Typically, the ICI term due to the SFO is very small and can be neglected.

6.6 IMPAIRMENT MITIGATION

We have seen that an OFDM system is sensitive to the phase-noise, the carrier-frequency offset, and the sampling frequency offset. NR is expected to be deployed at frequencies as high as 100 GHz, where these impairments would be a major bottleneck. In this section, we discuss mitigation schemes for these impairments assuming a slow time varying frequency-selective channel. In a fast varying channel, these impairments are typically corrected as part of the channel estimation and equalization process.

6.6.1 A PHASE-NOISE MITIGATION SCHEME

In order to focus on the PN impairment, we assume a perfect time-synchronization and a quasi-static multipath channel (with a channel length of L samples). The PN impaired signal at the receiver at the nth time sample can be expressed as

$$y[n] = \exp\left(j\phi[n]\right)\sum_{l=0}^{L-1} h_l x[n-l] + w[n], \tag{6.15}$$

where $x[n]$ represents the transmitted OFDM signal including the CP, h_l ($l = 0, \cdots, L\text{-}1$) is the lth tap of the channel impulse response (CIR), $w[n]$ denotes the additive white Gaussian noise (AWGN), and $\phi[n]$ denotes the PN.

Let $\mathbf{y} = \left[y[0]\ y[1]\ \cdots\ y[N-1]\right]^T$ be the received time-domain OFDM symbol (after CP removal), where N is the number of subcarriers. The PN is mitigated by the operation $\mathbf{\Phi y}$, where

$$\mathbf{\Phi} = \text{diag}\left\{\left[\exp\left(-j\hat{\phi}\,[0]\right)\ \cdots\ \exp\left(-j\hat{\phi}\,[N-1]\right)\right]^T\right\}, \tag{6.16}$$

with $\hat{\phi}$ denoting the estimate of ϕ. Hence, the task of the PN mitigation scheme is to essentially estimate the PN samples.

Let $\mathbf{H}_\text{p}$ be an $N_\text{p} \times N_\text{p}$ diagonal matrix consisting of the channel transfer functions (CTFs) at the N_p pilot subcarriers ($N_\text{p} \leq N$), $\mathbf{s}_\text{p}$ be a vector consisting of the N_p subcarriers, $\mathbf{D}$ be an $N_\text{p} \times N$ submatrix of the $N \times N$ discrete Fourier transform (DFT) matrix $\mathbf{F}$ (whose elements are given by $\exp(-j2\pi nk/N)/\sqrt{N}$ with $k,\ n \in \{0, \cdots, N-1\}$) corresponding to the N_p pilot subcarriers, and $\tilde{\mathbf{w}}$ be an $N \times 1$ vector consisting of the time-domain AWGNs. We have

$$\mathbf{D\Phi y} = \mathbf{H}_\text{p}\mathbf{s}_\text{p} + \mathbf{D\Phi}\tilde{\mathbf{w}}. \tag{6.17}$$

Let $\mathbf{Y} = \text{diag}(\mathbf{y})$ and $\mathbf{T}$ be an $N \times q$ transformation matrix, such that $\mathbf{\Phi} \approx \text{diag}(\mathbf{T}\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ consists of q unknowns or anchors ($q \leq N_\text{p}$), and it is given by

$$\boldsymbol{\alpha} = (\mathbf{DYT})^\dagger\,\mathbf{H}_\text{p}\mathbf{s}_\text{p} + (\mathbf{DYT})^\dagger\,\tilde{\mathbf{w}}. \tag{6.18}$$

The first term in the right hand side (RHS) of (6.18) is the least-square (LS) estimator of $\boldsymbol{\alpha}$, whereas the second term in the RHS of (6.18) is additive noise.

The physical meaning of $\boldsymbol{\alpha}$ depends on the type of the transformation matrix. If $\mathbf{T}$ is a linear interpolation matrix [5], the elements in $\boldsymbol{\alpha}$ are estimates of the inverse carrier PN $\exp(-j\phi)$ at the q anchors (time samples). These anchors are usually evenly distributed in the time-domain OFDM symbol.

Typically, the phase-noise power is concentrated in low frequencies. As an example, the power spectral density (PSD) of mmMAGIC phase-noise model at 82 GHz is shown in Fig. 6.24 [7]. As phase noise is dominant in lower frequencies, $\mathbf{T}$ can be a $N \times q$ submatrix of $\mathbf{F}$ corresponding to the q lowest spectral components. In this case, the elements in $\boldsymbol{\alpha}$ are the spectral components of $\exp(-j\phi)$.

The computational complexity of the PN mitigation scheme mainly depends on the pseudoinverse of an $N_\text{p} \times q$ matrix in (6.18), whose complexity increases linearly with N_p, yet cubically with q [2].

For convenience of the analysis, we study the performance of the PN mitigation/estimation in the preamble (where $N_\text{p} = N$). (The findings hold for the payload as well, where $N_\text{p} < N$.)

The LS estimator of the PN, i.e., the first term in the RHS of (6.18), contains the modeling error $\mathbf{\Phi} - \text{diag}(\mathbf{T}\boldsymbol{\alpha})$. Assuming perfect estimation of $\boldsymbol{\alpha}$ (by setting the second term in the RHS of (6.18), i.e., the additive noise, to zero), it is self-evident that the modeling error reduces to zero as q increases to N.

Now we examine the effect of the additive noise in (6.18) w.r.t. q. Let $\breve{\mathbf{w}} = \mathbf{F\Phi}\tilde{\mathbf{w}}$. Since $\mathbf{F\Phi}$ is a unitary matrix, $\breve{\mathbf{w}}$ and $\tilde{\mathbf{w}}$ have the same statistics. Thus, the additive noise in (6.18) can be equivalently written as

$$\mathbf{z} = (\mathbf{FYT})^\dagger\,\breve{\mathbf{w}} = (\mathbf{T})^\dagger\,(\mathbf{Y})^{-1}\,(\mathbf{F})^H\,\tilde{\mathbf{w}}. \tag{6.19}$$

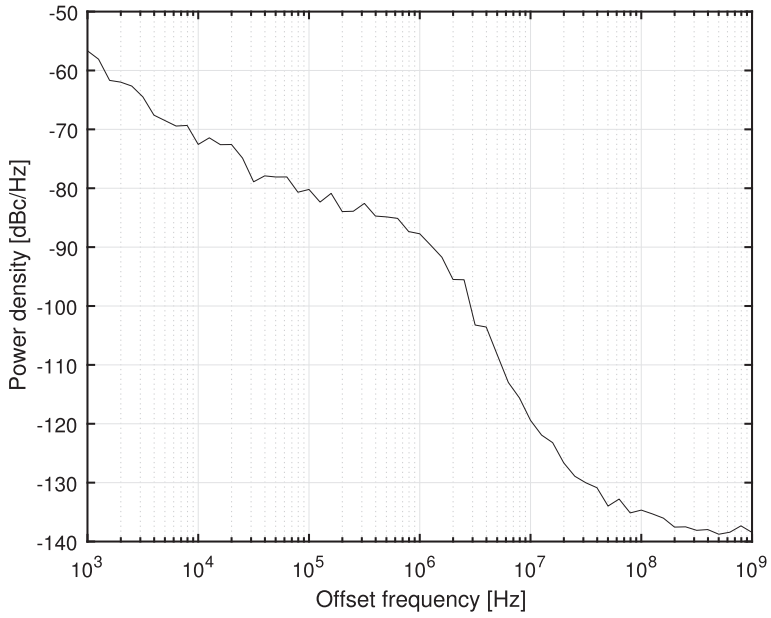


FIGURE 6.24

PSD of phase-noise.

For simplicity, we assume that $\mathbf{T}$ is the DFT transformation matrix. The power of the additive noise is given as

$$\begin{aligned}
\mathrm{E}\left[\mathbf{z}^H\mathbf{z}\right] &= \sigma_w^2 \mathrm{E}\left\{\mathrm{Tr}\left[(\mathbf{T})^\dagger (\mathbf{Y})^{-1}\left((\mathbf{Y})^{-1}\right)^*\left((\mathbf{T})^\dagger\right)^H\right]\right\} \\
&= \sigma_w^2 \mathrm{E}\left\{\mathrm{Tr}\left[\left(\mathbf{T}\mathbf{T}^H\right)^\dagger (\mathbf{Y})^{-1}\left((\mathbf{Y})^{-1}\right)^*\right]\right\} \\
&\geq \frac{q}{N}\frac{\sigma_w^2}{\sigma_h^2\sigma_s^2+\sigma_w^2} = \frac{q}{N}\frac{1}{\gamma_0+1},
\end{aligned} \tag{6.20}$$

where the last inequality follows from the Jensen's inequality and by denoting σ_h^2, σ_s^2, and σ_w^2 as variances of the CIR, subcarrier symbol, and AWGN, respectively; and $\gamma_0 = \sigma_h^2\sigma_s^2/\sigma_w^2$ is the signal-to-noise ratio (SNR). As can be seen, increasing q will increase the power of the additive noise in (6.18) for fixed N and SNR. This is because there are more anchors to the estimate as q increases. All in all, for the PN estimation, increasing q reduces the modeling error, yet increases the additive noises. As mentioned before, the computational complexity of the PN estimation increases cubically with increasing q. Therefore, q should not be unnecessarily large. For example, it has been shown that $q = 7$ is sufficient for mitigating the PN [4].

For simulations, we assume a DFT size of 512, including 32 scattered pilots. The remaining active subcarriers are loaded with 16- or 64-QAM. The CP length is set to be larger than the channel length. The sampling frequency is set so that the subcarrier spacing is 240 or 480 kHz. The carrier frequency

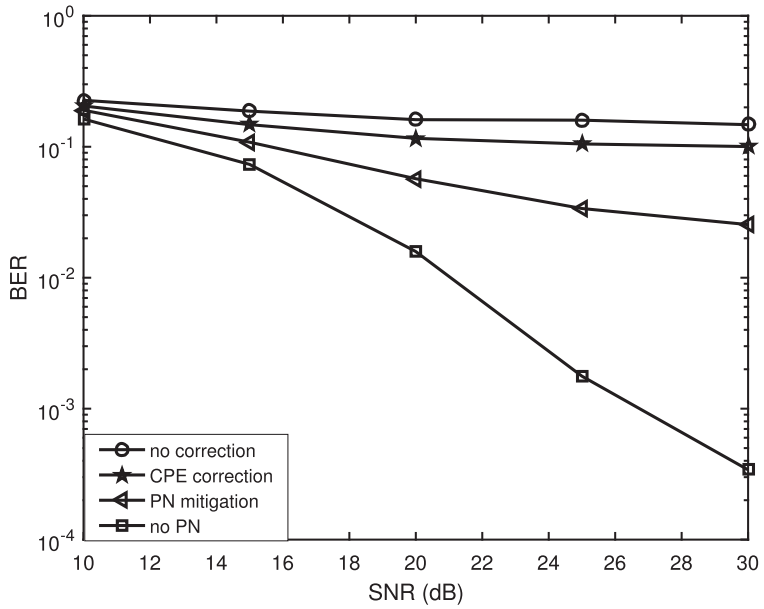


FIGURE 6.25

SER performance with 240 kHz subcarrier spacing and 64-QAM.

is set to fc = 82 GHz. The QuaDRiGa channel model [9] is used for channel emulation. Both the CPE correction [17] and the PN mitigation scheme with $q = 7$ are evaluated. Figs. 6.25–6.27 show the BER performances with/without phase-noise compensations with 16- and 64-QAM, and 240 and 480 kHz subcarrier spacing. It can be seen that the phase-noise mitigation scheme clearly outperforms the CPE correction (at the expense of increased complexity); and the corresponding BER performance improves as the subcarrier spacing increases from 240 to 480 kHz and as the modulation order decreases from 64-QAM to 16-QAM.

6.6.2 CFO AND SFO MITIGATION

In this section, we discuss estimation (and correction) of the CFO and the SFO. A scheme for mitigation of the SFO and the CFO (jointly with PN) is presented. This is an extension of the PN mitigation scheme presented in Section 6.6.1.

There exists an extensive literature on mitigation of the synchronization errors in OFDM systems. Thanks to the seminal work on the CFO estimation [16], [15], the (coarse) frequency synchronization can be readily achieved and our focus will be on the residual carrier-frequency offset estimation. The SFO mitigation is studied in [3], [19]. A pilot-based SFO estimator is proposed in [3], which estimates the SFO using scattered pilots in each OFDM symbol. The SFO can also be tracked using a delay-locked loop (DLL) [18]. A blind SFO estimator is proposed in [19], which estimates the SFO using current and previously received OFDM symbols. In general, the pilot-based SFO estimator outperforms the blind estimator and because the scattered pilots are needed for the PN estimation, the pilot-based SFO estimator is more popular in practical systems. To be robust to perturbations from an imperfect

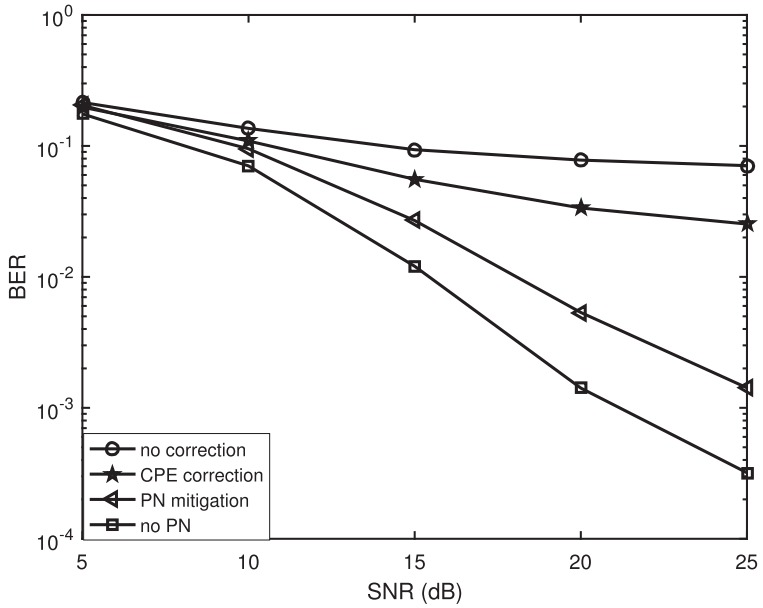


FIGURE 6.26

SER performance with 480 kHz subcarrier spacing and 64-QAM.

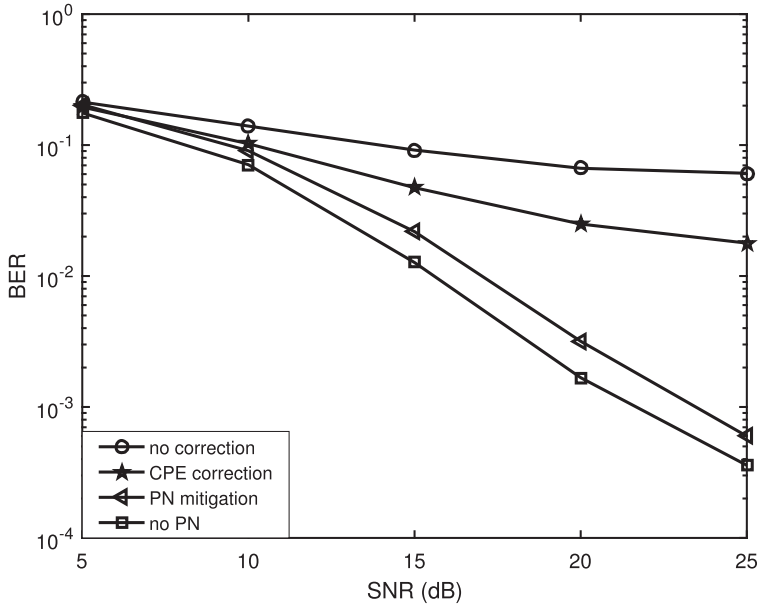


FIGURE 6.27

SER performance with 480 kHz subcarrier spacing and 16-QAM.

PN mitigation, ICI, and AWGN, the pilot-based SFO estimator can be combined with the Kalman filter to track the SFO. Unlike the DLL, the Kalman filter can also be used for channel prediction, which is necessary for the PN mitigation (prior to the SFO correction).

In the following, we present a mitigation scheme for the CFO and the SFO, assuming that there is a single clock and a single oscillator in the transmitter/receiver. For the case of multiple oscillators per transceiver, the interested reader may refer to [13], [6]. For simplicity and without loss of generality, we assume that coarse frequency synchronization has been achieved using the classical CFO estimation ([16,15]) with a residual CFO (RCFO) of ε. The residual CFO (RCFO) is normalized to the sampling frequency. Let T_{T} and T_{R} be the sampling intervals at the transmitter and at the receiver, respectively. The SFO can then be expressed as

$$\eta = \frac{T_{\mathrm{R}} - T_{\mathrm{T}}}{T_{\mathrm{T}}}. \tag{6.21}$$

The SFO is introduced to the transmit signal by sequentially sending the OFDM symbols to a buffer (which is longer than the duration of one OFDM symbol with CP), and sampling the buffered signal using the windowed sinc interpolation function with a sampling interval of $T_{\mathrm{T}}(1-\eta)$.

Consider the PN mitigation scheme presented in Section 6.6.1 with $\hat{\phi}[n]$ now denoting the estimate of the PN plus the RCFO at nth time sample,

$$\hat{\phi}[n] \approx \phi[n] + 2\pi\varepsilon n. \tag{6.22}$$

Once $\hat{\phi}[n]$ is estimated, the RCFO can be approximated by

$$\hat{\varepsilon} \approx \frac{1}{2\pi N} \sum_{n=0}^{N-2} \left(\hat{\phi}[n+1] - \hat{\phi}[n] \right). \tag{6.23}$$

The above expression assumes that the time derivative of the PN has zero mean, which holds for practical oscillators. After the PN mitigation, the RCFO becomes $\tilde{\varepsilon} \approx \varepsilon - \hat{\varepsilon}$.

The SFO causes a phase rotation (which varies linearly across subcarriers) and ICI. SFO is typically just a few ppm in practical systems; thus, the resulting ICI is negligible. For notational convenience, we drop the ICI term caused by the SFO in the following expressions. The received signal of the mth OFDM symbol at the kth subcarrier is given by

$$r_m[k] = J_m[0]\exp\left(j\frac{k\psi_m}{N}\right) H_m[k]s_m[k] + \sum_{l=0, l\neq k}^{N-1} J_m[k-l]\exp\left(j\frac{l\psi_m}{N}\right) H_m[l]s_m[l] + w_m[k], \tag{6.24}$$

where $s_m[k]$ and $r_m[k]$ are transmitted and received signals, $J_m[k]$ is the spectral component of the phase noise together with the RCFO (after PN mitigation), ψ_m is the rotation angle caused by the SFO together with the RCFO, $H_m[k]$ is the channel transfer function, and $w_m[k]$ is the AWGN. We have

$$J_m[k] = \frac{1}{N} \sum_{n=0}^{N-1} \exp\left(j\left(\tilde{\phi}_m[n] + 2\pi\tilde{\varepsilon}n\right)\right) \exp\left(-\frac{2\pi kn}{N}\right), \tag{6.25}$$

$$\psi_m = 2\pi m(N + N_g)\,(N\tilde{\varepsilon}/k + \eta)\,, \tag{6.26}$$

where $\tilde{\phi}_m[n]$ is the residual PN after the PN mitigation at the nth time sample of the mth OFDM symbol, and N_g denotes the CP length of the OFDM symbol. As can be seen, μ_m varies linearly over the OFDM symbol time. After an effective PN mitigation, $J_m(0)$ is close to one and the ICI term (i.e., the second term in (6.24)) is close to zero. Let $\tilde{H}_m[k] = J(0)H_m[k]$ and $\tilde{w}_m[k]$ be the sum of the ICI and AWGN terms, Eq. (6.24) can then be rewritten as

$$r_m[k] = \exp\left(j\frac{k\psi_m}{N}\right)\tilde{H}_m[k]s_m[k] + \tilde{w}_m[k]. \tag{6.27}$$

The (frequency-domain) phase rotation ψ_m must be corrected prior to the data detection. To be robust to perturbations of the residual PN and the AWGN, we apply the Kalman filter to track ψ_m. Let δ_m be the time derivative of ψ_m. The state-space model is given as

$$\mathbf{s}_{m+1} = \mathbf{M}\mathbf{s}_m + \mathbf{n}, \tag{6.28}$$

where $\mathbf{s}_m = [\psi_m\ \delta_m]^T$, $\mathbf{n}$ denotes the 2×1 state model noise vector, and

$$\mathbf{M} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

The covariance matrix of $\mathbf{n}$ is denoted $\mathbf{Q} = E[\mathbf{n}\mathbf{n}^H]$, where E is the expectation operator. The observation model is

$$\mu_m = \psi_m + \zeta, \tag{6.29}$$

where ζ denotes the observation noise with a variance of $\kappa = E[\zeta\zeta^*]$. By denoting the observation vector as $\mathbf{o} = [1\ 0]^T$; we can write

$$\mu_m = \mathbf{o}^T\mathbf{s}_m + \zeta. \tag{6.30}$$

The Kalman filter algorithm is now given by:

Initialize:

$$\begin{aligned} \mathbf{s}_{0|0} &= \begin{bmatrix} 0 & 0 \end{bmatrix}^T \\ \mathbf{P}_{0|0} &= \mathbf{I}_2. \end{aligned} \tag{6.31}$$

Iterate:

predict:

$$\begin{aligned} \mathbf{s}_{m|m-1} &= \mathbf{M}\mathbf{s}_{m-1|m-1} \\ \mathbf{P}_{m|m-1} &= \mathbf{M}\mathbf{P}_{m-1|m-1}\mathbf{M}^T + \mathbf{Q} \end{aligned} \tag{6.32}$$

update:

$$\begin{aligned}
\alpha_m &= \mu_m - \mathbf{o}^T \mathbf{s}_{m|m-1} \\
\mathbf{k}_m &= \frac{\mathbf{P}_{m|m-1}\mathbf{o}}{\mathbf{o}^T \mathbf{P}_{m|m-1}\mathbf{o}} \\
\mathbf{s}_{m|m} &= \mathbf{s}_{m|m-1} + \mathbf{k}_m \alpha_m \\
\mathbf{P}_{m|m} &= (\mathbf{I}_2 - \alpha_m \mathbf{k}_m)\, \mathbf{P}_{m|m-1}.
\end{aligned} \tag{6.33}$$

At the mth OFDM symbol time, the phase rotation has to be estimated based on the scattered pilots in the mth OFDM symbol. Let k_p $(p = 1,\ ...,\ N_\mathrm{p})$ be the indices of the N_p pilots in the mth OFDM symbol, and denote

$$\vartheta_m[k] = \angle \left(\frac{\left(\tilde{H}_m[k] s_m[k]\right)^* r_m[k]}{\left(\tilde{H}_m[k] s_m[k]\right)^* \left(\tilde{H}_m[k] s_m[k]\right)} \right), \tag{6.34}$$

where $\angle(\cdot)$ denotes the angle of its argument, and the superscript * denotes the complex conjugate operation. The phase rotation can be estimated as

$$\mu_m = \frac{N}{N_\mathrm{p} - 1} \sum_{p=1}^{N_\mathrm{p}-1} \frac{\vartheta_m[k_p] - \vartheta_m[k_{p+1}]}{k_p - k_{p+1}}. \tag{6.35}$$

When $|\mu_m|$ is larger than 2π, the synchronization module shifts the signal by one sample accordingly; otherwise, the SFO is corrected by $\exp(-jk\mu_m/N)\, \mathbf{y}_m[k]$. Note that the SFO correction also compensates for the RCFO $\tilde{\varepsilon}$. The PN mitigation and the SFO correction are applied to both the training signal and the payload. In the payload, the estimated channel is used together with the Kalman filter for the channel prediction. Specifically, we predict the phase rotation for the next OFDM symbol:

$$\mu_{m+1}[k] = \mathbf{o}^T \mathbf{M} \mathbf{s}_{m|m}. \tag{6.36}$$

The channel prediction for the next OFDM symbol is

$$\hat{H}_{m+1}[k] = \exp\left(\frac{jk\mu_{m+1}}{N}\right) \hat{H}_m[k], \tag{6.37}$$

which is used for the PN mitigation in the next OFDM symbol. We assume that the OFDM payload has 512 subcarriers, including 354 active subcarriers, 32 pilot subcarriers, and 158 guard band subcarriers (79 null subcarriers in the beginning and 79 at the end of each OFDM symbol). The guard band helps to reduce the out-of-band emission and eliminates the distortions in the interpolation and the decimation process. The CP length is set to be larger than the channel length. We further assume that the system is operating at 60 GHz with 256-QAM and sampling frequency 100 MHz. For tracking the frequency-domain phase rotation (due to the SFO and the RCFO) using the Kalman filter, we assume the covariance matrix of the state model noises to be $\mathbf{Q} = 10^{-15}\mathbf{M}\mathbf{M}^T$ and the variance of

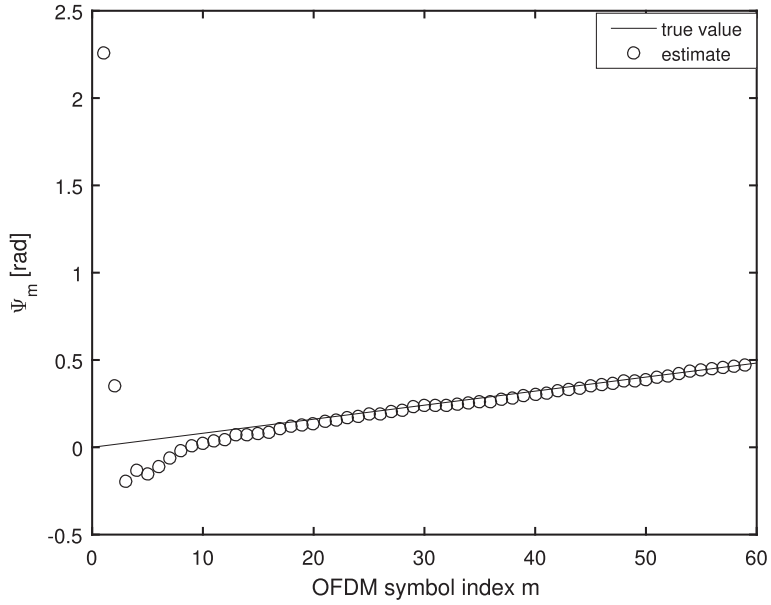


FIGURE 6.28

Example of the SFO phase rotation and its estimate as a function of the OFDM symbol index m.

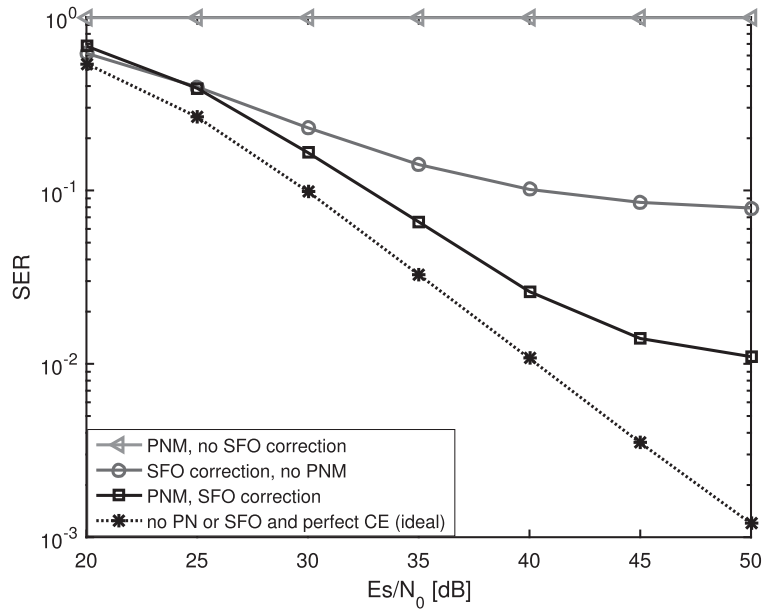


FIGURE 6.29

SER performances of the uncoded MIMO-OFDM system with/without the PN mitigation or the SFO correction in an urban microcell scenario at 60 GHz.

the observation noise to be $\kappa = 0.01$. Note that the small covariance matrix $\mathbf{Q}$ is justified by the time-invariant state model. The observation noise ζ is actually the estimation error of the phase rotation estimator, whose mean square error (MSE) is nontrivial to derive. However, it is found that the error rate performance of the system is insensitive to the value of κ (i.e., the variance of ζ) ranging from 0.001 to 0.1. Fig. 6.28 shows an example of the estimate of the phase rotation ψ_m and its true values as functions of the OFDM symbol index m. As can be seen, the estimates converge after the first eight OFDM symbols. Note that, apart from being robust to perturbations, e.g., of the PN estimation error, the Kalman filter is also used for the channel prediction for the PN mitigation.

For evaluating the error rate performance of the MIMO-OFDM link in a multipath fading channel, we resort to the QuaDRiGa channel model [9]. In total 200 drops are generated. For each channel drop, 10 preamble symbols and 50 payload symbols are transmitted. Fig. 6.29 shows the symbol error rate (SER) performance of the uncoded MIMO-OFDM system in the presence of the PN, the SFO, and the RCFO with/without the PN mitigation (PNM) and the SFO correction. For comparison, we also plot SERs for the ideal case without the PN, the SFO, the RCFO, and with a perfect channel estimation. As can be seen, the PN and SFO must be jointly compensated for in order to have an acceptable SER performance.

REFERENCES

[1] E. Bala, J. Li, R. Yang, Shaping spectral leakage: a novel low-complexity transceiver architecture for cognitive radio, IEEE Vehicular Technology Magazine 8 (3) (2013, Sept.) 38–46.

[2] P. Bürgisser, M. Clausen, A. Shokrollahi, Algebraic Complexity Theory, Springer-Verlag, Berlin/Heidelberg, 1997.

[3] D.C. Chang, Effect and compensation of symbol timing offset in ofdm systems with channel interpolation, IEEE Transactions on Broadcasting (ISSN 0018-9316) 54 (4) (2008, Dec.) 761–770, https://doi.org/10.1109/TBC.2008.2002339.

[4] X. Chen, OFDM based multi-node transmission in the presence of phase noises for small cell backhaul, IEEE Communications Letters (ISSN 1089-7798) 21 (5) (2017, May) 1207–1210, https://doi.org/10.1109/LCOMM.2017.2655509.

[5] X. Chen, A. Wolfgang, Phase noise mitigation in ofdm-based backhaul in the presence of channel estimation and synchronization errors, in: 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring), 2016, May, pp. 1–5.

[6] X. Chen, C. Fang, Y. Zou, A. Wolfgang, T. Svensson, Beamforming mimo-ofdm systems in the presence of phase noises at millimeter-wave frequencies, in: 2017 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), 2017, March, pp. 1–6.

[7] D5.1, Initial multi-node and antenna transmitter and receiver architectures and schemes, Millimetre-Wave Based Mobile Radio Access Network for Fifth Generation Integrated Communications (mmMAGIC), 2016.

[8] G. Huang, A. Nix, S. Armour, Impact of radio resource allocation and pulse shaping on PAPR of SC-FDMA signals, in: Proc. IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2007, Sept., pp. 1–5.

[9] S. Jaeckel, L. Raschkowski, K. Börner, L. Thiele, Quadriga: a 3-d multi-cell channel model with time evolution for enabling virtual field trials, IEEE Transactions on Antennas and Propagation (ISSN 0018-926X) 62 (6) (2014, June) 3242–3256, https://doi.org/10.1109/TAP.2014.2310220.

[10] D.W. Lim, S.J. Heo, J.S. No, An overview of peak-to-average power ratio reduction, Journal of Communications and Networks 11 (3) (2009, June) 229–239.

[11] A. Mehrotra, Noise analysis of phase-locked loops, in: IEEE/ACM International Conference on Computer Aided Design. ICCAD – 2000. IEEE/ACM Digest of Technical Papers (Cat. No.00CH37140), 2000, Nov., pp. 277–282.

[12] S.H. Muller-Weinfurtner, Optimum Nyquist windowing in OFDM receivers, IEEE Transactions on Communications 49 (3) (2001, Mar.) 417–420.

[13] A. Puglielli, G. LaCaille, A.M. Niknejad, G. Wright, B. Nikolić, E. Alon, Phase noise scaling and tracking in OFDM multi-user beamforming arrays, in: 2016, IEEE International Conference on Communications (ICC), 2016, May, pp. 1–6.

[14] R1-162960, Summary of email discussion on link level channel model, 3GPP TSG RAN WG1 Meeting 84b, 2016, April.

[15] T.M. Schmidl, D.C. Cox, Robust frequency and timing synchronization for ofdm, IEEE Transactions on Communications (ISSN 0090-6778) 45 (12) (1997, Dec.) 1613–1621, https://doi.org/10.1109/26.650240.
[16] J.J. van de Beek, M. Sandell, P.O. Borjesson, Ml estimation of time and frequency offset in OFDM systems, IEEE Transactions on Signal Processing (ISSN 1053-587X) 45 (7) (1997, July) 1800–1805, https://doi.org/10.1109/78.599949.
[17] S. Wu, Y. Bar-Ness, OFDM systems in the presence of phase noise: consequences and solutions, IEEE Transactions on Communications (ISSN 0090-6778) 52 (11) (2004, Nov.) 1988–1996, https://doi.org/10.1109/TCOMM.2004.836441.
[18] B. Yang, K.B. Letaief, R.S. Cheng, Z. Cao, Timing recovery for OFDM transmission, IEEE Journal on Selected Areas in Communications (ISSN 0733-8716) 18 (11) (2000, Nov.) 2278–2291, https://doi.org/10.1109/49.895033.
[19] Y.H. You, S.T. Kim, K.T. Lee, H.K. Song, An improved sampling frequency offset estimator for ofdm-based digital radio mondiale systems, IEEE Transactions on Broadcasting (ISSN 0018-9316) 54 (2) (2008, June) 283–286, https://doi.org/10.1109/TBC.2008.915763.
[20] A.A. Zaidi, R. Baldemair, H. Tullberg, H. Bjorkegren, L. Sundstrom, J. Medbo, C. Kilinc, I.D. Silva, Waveform and numerology to support 5G services and requirements, IEEE Communications Magazine (ISSN 0163-6804) 54 (11) (2016, November) 90–98, https://doi.org/10.1109/MCOM.2016.1600336CM.
[21] J. Zhang, C. Huang, G. Liu, P. Zhang, Comparison of the link level performance between OFDMA and SC-FDMA, in: Communications and Networking in China (ChinaCom), 2006, Oct., pp. 1–6.

CHAPTER

MULTIANTENNA TECHNIQUES 7

Multiantenna techniques can be defined by the use of multiple antennas at the transmitter and/or receiver in combination with signal processing. A communication system with multiple antennas at both the transmitter and the receiver is often referred to as a MIMO system. These techniques can be used to improve the system performance in terms of capacity, coverage, data rates, and link reliability. Fundamentally, multiple antennas can provide:

- Diversity. Diversity gives robustness against a fading radio channel and improves the link reliability. Multiple antennas can be used in different ways to provide diversity, e.g., space, pattern, polarization, and delay diversity.
- Array gain. By coherently combining the signals from/to multiple antennas, so that the desired signal adds constructively, a spatial processing gain can be achieved that increases the SNR. In a LoS channel, array gain is obtained by directivity but array gain can also be obtained in rich scattering channels by proper antenna combining.
- Interference suppression. Multiple antennas cannot only be used to enhance the desired signal but they can also be combined so that undesired signals are suppressed, thereby also improving the signal-to-interference-ratio (SIR).
- Spatial multiplexing. By transmitting multiple data streams over several antennas using the same time-frequency radio resource, spectral efficiency can be improved. The multiple data streams can be transmitted to a single UE, often called single-user MIMO (SU-MIMO), or multiple UEs, often called MU-MIMO.

Multiple antennas have been used in global wireless communication systems such as GSM, WCDMA/HSPA, and LTE. Initially, only simple antenna diversity techniques were used. More advanced multiantenna techniques such as MIMO were introduced in HSPA Release 7 and developed further in LTE. While multiantenna techniques have been useful for improving performance in current and previous generations of the standards, in 5G NR they have a more fundamental role to play in the system design.

This chapter gives an overview of multiantenna techniques in cellular wireless communication and their particular use in NR. The NR-specific features and functionalities described in this chapter are according to the first NR release (Release 15) which was finalized in 2018. NR is continuously evolving and new features will be added and existing features will be enhanced in future releases of the specifications.

The chapter is organized as follows. In Section 7.1 we discuss the role of multiantenna techniques in NR, for both low and high frequency bands. The fundamental theory of the multiantenna techniques relevant for NR is provided in Section 7.2 in order to give a better understanding of the particular features adopted in the NR specifications. The NR-specific multiantenna techniques and features are then described in Section 7.3. Finally, the effectiveness of the discussed techniques is illustrated by some experimental examples in Section 7.4.

5G Physical Layer. https://doi.org/10.1016/B978-0-12-814578-4.00012-6
Copyright © 2018 Elsevier Ltd. All rights reserved.

7.1 THE ROLE OF MULTIANTENNA TECHNIQUES IN NR

This section gives a brief overview of the role of multiantenna techniques in NR. More details on specific features are provided in later sections. NR has been designed for millimeter-wave spectrum in addition to traditional cellular frequency bands at lower frequencies. The motivation for having multiple antennas and the techniques used are different in the low and high frequency bands. Some of these aspects are discussed in the following. More attention is paid to the high frequency bands, since this is where multiantenna techniques have a more fundamental impact on the system design and where most new features have been developed. Clearly, there is no sharp border between low and high frequencies in general. However, 3GPP NR has defined two frequency ranges, FR1 and FR2, where FR1 is between 450 MHz and 6 GHz and FR2 is between 24.25 GHz and 52.6 GHz[1] [5]. Therefore, in this chapter, low frequencies will mean carrier frequencies $\leq$ 6 GHz and high frequencies will mean carrier frequencies $\geq$ 24.25 GHz. High frequencies will sometimes also be referred to as millimeter-wave frequencies

7.1.1 LOW FREQUENCIES

In the low frequency bands, spectrum is congested. To meet the never-ceasing quest for higher data rates, a higher spectral efficiency is needed. This can be achieved by advanced multiantenna techniques such as spatial multiplexing and interference suppression. The former is attained by reusing radio resources in an efficient manner and the latter by ultimately enabling a multicell, multiuser system to be limited by thermal noise and not by interference.

At low frequencies the physical size may limit the number of antenna elements in an array that is practical to have since the antenna element area is proportional to the wavelength squared; see Chapter 3. However, advances in active array antenna technology have made it feasible to have digital control over more of the antenna elements in the array. This can be used to exploit more details in the spatial domain in order to increase performance.

For low frequencies, multiantenna techniques for NR are mainly refinements and evolution of multiantenna techniques that have been used in LTE. Some of the enhancements include improved support for reciprocity-based operation and more detailed feedback of CSI to achieve a higher spectral efficiency with MU-MIMO transmission.

7.1.2 HIGH FREQUENCIES

At high frequencies, the spectral efficiency is less crucial since there is plenty of spectrum available. Instead, obtaining coverage is the main challenge as substantially higher transmission losses may occur due to smaller antenna apertures and in some cases also higher attenuations, as explained in Chapter 3. A large bandwidth can exacerbate this further due to increased thermal noise power in the receivers. Compensating this with higher transmission power might not be possible due to limitations in millimeter-wave hardware design and to current regulations on transmitted power being stricter at frequencies above 6 GHz [11]. On the other hand, increasing the carrier frequency also means that, for a given physical size of an antenna, it becomes more directive. This may, depending on the directional

[1] The gap between 6 GHz and 24.25 GHz is due to the fact that no spectrum allocations have been identified for NR in this range.

properties of the channel, compensate for the increase of transmission loss with frequency and even be turned into a gain, as explained in Chapter 3. To make use of this directivity, dynamic and user-specific beam-forming is needed, since the directions to the users in a wireless access network are not known a priori and are dynamically changing.

In free-space propagation, the transmission loss increases with frequency if the antenna gain is assumed to be constant at both ends of the link; see Chapter 3. This is an effect of the fact that the antenna effective area is proportional to the wavelength squared, as explained in Chapter 3. If instead the antenna area is assumed to be constant at one end of the link, the free-space transmission loss is frequency independent. If the antenna area is assumed to be constant at both ends of the link, the free-space transmission loss will actually decrease with frequency since the antenna gain increases with frequency if the antenna area is constant. However, to make use of the increased antenna gain the transmit (Tx) and receive (Rx) beams must be aligned. For a point-to-point radio link this can be achieved by mechanically aligning the Tx and Rx antennas. However, in a mobile communication system with moving users this solution is not practically feasible. Instead, beam-forming and beam tracking using array antennas are needed to dynamically adjust the directions of the Tx and Rx beams. If the frequency is increased, more antenna elements can be accommodated by an antenna array with a given physical size, since the individual antenna elements become smaller. Therefore the potential beam-forming gain increases with frequency for a given physical size of the array.

Clearly, a mobile communication system cannot rely only on free-space propagation since many users have NLoS and/or are located indoors. Therefore, other propagation characteristics such as diffraction, reflection, scattering, and penetration are important to consider. These different propagation mechanisms have a varying degree of frequency dependence; see Chapter 3 for more details. However, in many NLoS scenarios there is a substantial increase of transmission loss with frequency making it more challenging to maintain an adequate link budget at high frequencies. To what extent beam-forming can compensate for the increased propagation loss depends on the scenario and also on how the beam-forming is implemented. Both theoretical and experimental examples of this issue are given later in this chapter.

The adverse propagation conditions and current hardware technology at millimeter-wave frequencies have a fundamental impact on the NR system design. To ensure sufficient coverage, in the millimeter-wave spectrum NR has a beam-centric design in which not only data transmissions can be beam-formed but also control and broadcast signals. This is different from previous generations of cellular systems in which typically only data transmissions are beam-formed.[2] Furthermore, support for beam-forming also in the UE has been introduced in NR in order to increase the potential beam-forming gain even further. UE beam-forming is possible in millimeter-wave bands since more antenna elements can be fit into the limited form factor of a UE. Due to hardware constraints, analog beam-forming will be a common implementation in millimeter-wave bands, particularly for hand-held devices. Therefore, support for analog beam-forming procedures has been included in the NR specifications.

Another difference between low and high frequency bands is that spectrum allocations in millimeter-wave bands are foreseen to mainly be unpaired. This has impact on the duplexing method

[2]Exceptions exist, e.g., the enhanced physical downlink control channel (EPDCCH) introduced in LTE Release 11 is a control channel that can be beam-formed.

used and thereby also on how different multiantenna techniques can operate. Frequency division duplex (FDD) operation is used in paired frequency bands where different frequency ranges are assigned for downlink and uplink, while TDD is used in unpaired bands where a single frequency range is shared between downlink and uplink. This has significant impact on multiantenna techniques, since the propagation channel can be assumed to be reciprocal under TDD operation, i.e., the downlink channel state is identical to the uplink channel state. In FDD operation, the downlink and uplink will typically experience independent fast fading due to the frequency difference between the uplink and downlink carriers. Advanced multiantenna transmission techniques often rely on detailed channel state information at the transmitter (CSIT). If reciprocity holds, this can be obtained from uplink measurements while extensive pilot and feedback signaling may be needed in the case that it does not. Therefore, reciprocity-based multiantenna transmission techniques can benefit from TDD operation, especially for antenna arrays with many elements where the signaling overhead may become prohibitive. Note that not only the propagation channel needs to be reciprocal, but also the multiantenna transceivers, which may require calibration of the transceivers [50].

7.2 MULTIANTENNA FUNDAMENTALS

In this section we provide some fundamental theory of multiantenna techniques relevant for NR. We try to keep the presentation general, thus being agnostic to any particular wireless communication standard. However, when deemed relevant, we at times point out relations to current LTE and NR specifications. Details on particular techniques adopted in the NR specifications are deferred to Section 7.3.

7.2.1 BEAM-FORMING, PRECODING, AND DIVERSITY

Beam-forming, precoding,[3] and diversity are techniques to coherently combine multiple antenna elements in an antenna array; at the transmitter, receiver or both.[4] By doing so, two types of gain can be achieved:

- Array gain. Array gain is the increase in the average SNR obtained by combining multiple antenna elements compared to a single element.
- Diversity gain. Diversity techniques are used to reduce the impact of fading by combining antenna elements that experience different fading. Diversity performance can be characterized by diversity order, which is the number of independently fading antenna elements.

The difference between array gain and diversity gain is that array gain gives an increase in the average SNR, while diversity gain makes the probability density function of the instantaneous SNR more concentrated around its average value. In some cases, array and diversity gain can be achieved simultaneously, while in other cases only one of the gains is achieved. This is described later in this section.

[3]Precoding is a transmission technique. For lack of better terminology we simply call it antenna combining when performed in the receiver.

[4]Diversity also includes other techniques than antenna combining such as antenna selection, frequency diversity, etc.

One of the most simple, intuitive and often used multiantenna techniques is beam-forming. Beam-forming is a technique to focus the transmitted or received radio energy in a particular direction using an array of antenna elements. This is achieved by applying a progressive time delay to the antenna elements so that the signals from the different elements add constructively in a desired direction. By controlling the progressive time delay, a beam can be steered in a desired direction. To steer a beam in a certain direction the time delays should compensate the propagation delay between the antenna elements of a plane wave impinging upon the array from that direction. For a narrowband system, a time delay can be approximated by a phase shift. Beam-forming is implemented by multiplying the signals on the antenna elements by complex beam-forming weights. The phase of the beam-forming weights determine the beam direction and the amplitude can be used to control beam width and sidelobe level.

The notion of beam-forming is coupled to free-space propagation of a single plane wave.[5] Under such conditions, beam-forming is the optimal transmission/reception scheme in the sense that it maximizes the SNR if the noise is spatially white. An antenna array with N elements can achieve an array gain that is equal to N. In wireless communication, beam-forming is a suitable technique for scenarios in which there is one dominating propagation path, e.g., when there is LoS or a strong specular reflection. The transmitter can apply Tx beam-forming to focus the transmitted energy in the dominating angle of departure and the receiver can apply Rx beam-forming to obtain a focused reception in the dominating angle of arrival.

The Tx and Rx array gains in a "pure" LoS channel, i.e., a fully correlated channel with no multipath propagation, are multiplicative so that the composite link array gain for a transmitter with N_T elements and a receiver with N_R elements is $N_T N_R$. Hence, full Tx and Rx array gain can be obtained simultaneously in a fully correlated channel. Under these circumstances there is no fast or frequency-selective fading. Furthermore, in most scenarios, the LoS direction varies slowly. Therefore, in this case, the beam-forming weights can be updated on a slow time basis and the same weights can be used over the entire bandwidth.[6]

However, most cellular deployments are characterized by multipath propagation. This means that the communication between a transmitter and receiver is not conveyed by a single plane wave, rather a superposition of multiple plane waves, or channel rays. This superposition is the cause of fast fading in wireless communications; see Chapter 3. The rays have different angles of arrival/departure as well as different amplitudes and phases. This means that it is no longer optimal to transmit/receive in a single direction by applying a simple progressive phase shift. Instead, the optimal approach is to utilize the different propagation paths in the channel. By proper coherent combining of the antenna elements, energy can actually be focused in space (within a limited resolution) rather than in direction. In effect, the different propagation paths are then aligned so that they add constructively at the receiver. Precoding can also be applied to antenna elements having different polarizations to match the polarization properties of the channel.

[5]The plane wave assumption implies that the receiver is in the far field of the Tx antenna. One rule of thumb for the far field is a minimum distance of $2L^2/\lambda$, where L is the largest array dimension and λ is the carrier wavelength [30]. This is fulfilled in a mobile communication network.

[6]If not too large to violate the narrowband assumption. This depends both on the relative bandwidth and the size of the antenna array in wavelengths. Different rules of thumb can be derived depending on the criterion. One example is that the maximum relative bandwidth in per cent is equal to the array beam width in degrees [30].

This more general amplitude and phase combining of antenna elements is commonly referred to as precoding when applied at the transmitter, or antenna combining when applied at the receiver. Also precoding is implemented by applying complex weights to the antenna elements. Although there is no strict distinction between beam-forming and precoding, beam-forming may be viewed as a special case of precoding for correlated channels. In a rich scattering channel, the fading correlation between different antenna elements will be low.[7] Precoding can therefore also provide diversity gain in addition to array gain. This requires more detailed channel knowledge, and the precoding weights need to be updated more frequently to follow the fast fading. Furthermore, fast fading is usually frequency selective so that different precoding weights may be needed in different parts of the scheduled bandwidth.

To design an optimal precoder, the complex channel coefficients between the transmitter and receiver antenna elements need to be known. How to acquire this information is a challenging task, which is described in more detail in Section 7.2.6. For now, assume that the complex channels between all Tx/Rx antenna pairs are known to both the transmitter and the receiver. If the transmitter emits a symbol, s, precoded with an $N_T \times 1$ complex weight vector $\mathbf{w}_T$ using an array with N_T elements, the signal at a receiver having an array with N_R elements can be modeled by

$$\mathbf{y} = \sqrt{P}\mathbf{H}\mathbf{w}_T s + \mathbf{n}, \tag{7.1}$$

where $\mathbf{y}$ is an $N_R \times 1$ vector containing the signal samples from the Rx antennas, P is the transmitted power per antenna, $\mathbf{H}$ the $N_R \times N_T$ complex channel matrix, and $\mathbf{n}$ is an $N_R \times 1$ vector modeling additive noise, which is assumed to be spatially white. The receiver can apply a $1 \times N_R$ antenna combining weight vector $\mathbf{w}_R$ to produce the complex scalar output

$$z = \mathbf{w}_R\mathbf{y} = \sqrt{P}\mathbf{w}_R\mathbf{H}\mathbf{w}_T s + \mathbf{w}_R\mathbf{n}. \tag{7.2}$$

First, assume that the transmitter has a single antenna element and the receiver has N_R antenna elements so that $\mathbf{H} = \mathbf{h}$, where $\mathbf{h}$ is the $N_R \times 1$ channel vector. The Rx combining vector that maximizes the SNR after the combiner when the noise is spatially white is easily shown[8] to be

$$\mathbf{w}_R = \frac{\mathbf{h}^H}{\|\mathbf{h}\|_F} \tag{7.3}$$

when $\mathbf{w}_R$ is constrained to fulfill $\|\mathbf{w}_R\|_F = 1$. Here, $\mathbf{h}^H$ denotes the complex conjugate transpose of $\mathbf{h}$ and $\|\cdot\|_F$ denotes the Frobenius norm. This solution is commonly referred to as maximum ratio combining (MRC). It is also easy to show that MRC achieves full array gain equal to N_R, regardless of the channel correlation.

Second, assume that the receiver has a single antenna and the transmitter has N_T antennas. The precoding vector that maximizes the SNR at the receiver is given by

$$\mathbf{w}_T = \frac{\mathbf{h}^H}{\|\mathbf{h}\|_F} \tag{7.4}$$

[7] The fading correlation is determined by the distance between antenna elements and the channel angular spread.

[8] This follows from the Cauchy–Schwarz inequality.

when $\mathbf{w}_T$ is constrained to fulfill $\|\mathbf{w}_T\|_F = 1$ and where $\mathbf{h}$ is the channel vector, presently a $1 \times N_T$ vector. The precoder in (7.4) is often called an MRT precoder. The MRT precoder achieves full array gain, N_T, regardless of the channel correlation. A physical explanation of this is that the precoder makes the signals scattered by the channel arrive in phase and add coherently at the receiver. It is important to realize that the power constraint assumed in the derivation is on the total transmitted power summed over all antenna elements, not on the power per antenna element. In general, the MRT solution gives a precoder with non-constant modulus weights, i.e., the weights for different antenna elements have different amplitudes. An active array typically has a PA per antenna element so that power cannot be shared between antenna elements. Therefore, a per-antenna power constraint would be a more realistic assumption for an active array antenna. Maximizing the SNR under a per-antenna power constraint is achieved by simply keeping the phase of the MRT precoder and setting all amplitudes equal [53]; sometimes this is called an equal gain transmission (EGT) precoder. The average SNR loss for an EGT precoder compared to MRT assuming an IID Rayleigh fading channel is $N_T/(1+(N_T-1)\pi/4)$, which converges to $4/\pi$ or roughly 1 dB for large N_T [44]. Note that this loss is under the assumption of the same total Tx power for the two precoders. If an MRT precoder with non-constant modulus weights is applied to an array with a PA per antenna element, not all PAs would transmit with full power. This would lead to a reduction in the total transmitted power and a corresponding reduction in SNR for the MRT precoder.

To illustrate the difference between precoding in a LoS and NLoS scenario, Fig. 7.1 shows an example of azimuth radiation patterns for these two cases. The left plot shows the radiation pattern of a uniform linear array (ULA) with 16 antenna elements when the optimal MRT precoding weights have been applied to the array for a scenario in which there is LoS to the UE. The right plot shows a corresponding pattern when the UE is in NLoS, based on one realization of the ITU urban macro channel model in [22]. In the LoS case, there is a single, narrow beam pointing in the direction to the UE. In the NLoS case, energy is instead transmitted in several different directions in order to utilize different propagation paths in the environment. The radiation pattern of the MRT precoder matches the azimuth power spectrum of the channel. In the LoS case, MRT precoding is equivalent to beam-forming according to our previous "definition". In the NLoS case, we prefer to use the term precoding since it is hard to distinguish a well-defined main beam from such a radiation pattern.[9] Note that the maximum directivity of the radiation pattern is lower in the NLoS case due to the multipath propagation. Nevertheless, the array gain is the same for both cases since the "multiple beams" in the NLoS case are combined coherently at the receiver.

Third, and finally, assume that the transmitter has N_T antenna elements and the receiver has N_R antenna elements. The precoder and combiner that maximize the SNR at the receiver output under a sum-power constraint is given by the principal right- and left-singular vector of $\mathbf{H}$, respectively [32], i.e.,

$$\mathbf{w}_T = \mathbf{v}_1, \ \mathbf{w}_R = \mathbf{u}_1^H \tag{7.5}$$

[9] According to IEEE, a beam is defined as "The major lobe of the radiation pattern of an antenna" and a major lobe is defined as "The radiation lobe containing the direction of maximum radiation" [21].

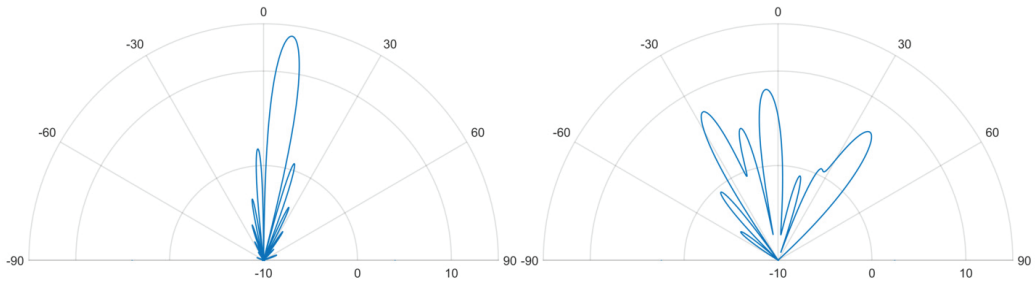


FIGURE 7.1

Radiation patterns for MRT precoding. Sample radiation pattern for a ULA with 16 elements performing MRT precoding. Left: There is LoS to the UE. Right: The UE is in NLoS.

where $\mathbf{v}_1$ is the first[10] column of $\mathbf{V}$ and $\mathbf{u}_1$ is the first column of $\mathbf{U}$ in the singular value decomposition (SVD) of $\mathbf{H}$

$$\mathbf{H} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H. \tag{7.6}$$

The composite link array gain of this scheme, sometimes referred to as dominant eigenmode transmission, is equal to the expectation of the maximum eigenvalue of $\mathbf{H}\mathbf{H}^H$ [32]. In this case, the composite link array gain depends on the channel correlation. In a channel with full Tx and Rx correlation, e.g., a pure LoS channel, the composite link array gain is $N_R N_T$ so that full Rx and Tx gain is achieved simultaneously. In a scattering channel, however, the composite link array gain will be lower, since a single precoder cannot maximize the power at all Rx antenna elements simultaneously. It can be shown that, for an uncorrelated channel, the composite link array gain is upper bounded by $(\sqrt{N_T} + \sqrt{N_R})^2$ for large arrays [8]. Hence, full Tx array gain *or* full Rx array gain can be obtained in a scattering channel, but not the two simultaneously. This can only be achieved in a fully correlated channel. Similar to the previous case of a single-antenna receiver, all amplitudes in $\mathbf{w}_T$ can be set to the same value in order to fully utilize all PAs also when the receiver has multiple antennas.[11] It has been shown that the incurred SNR loss by doing this is not more than 1 dB for an IID Rayleigh channel, i.e., not more than when we have a single Rx antenna [45].

MRC and MRT were derived under the assumption of spatially white noise. This is a reasonable assumption for thermal noise but in many cases the dominating impairment is caused by interference. Interference is typically not spatially white, since it usually comes from a particular direction. The antenna elements can then be combined with weights so that the array beam pattern has nulls in the interference directions. Theoretically, $N - 1$ interference directions can be suppressed with an array having N antenna elements. Interference can be suppressed in the receiver, see Section 7.2.2.3, or by the transmitter, see Section 7.2.2.2.

Besides array gain, precoding and antenna combining also gives diversity gain in a scattering channel since the signals transmitted/received on spatially separated antennas or antennas with different

[10] Assuming the singular values in $\boldsymbol{\Sigma}$ have been ordered in descending order.

[11] The Rx combining vector should then be modified to $\mathbf{w}_R = \tilde{\mathbf{h}}/\|\tilde{\mathbf{h}}\|_F$, where $\tilde{\mathbf{h}} = \mathbf{H}\mathbf{w}_{\text{EGT}}$, and $\mathbf{w}_{\text{EGT}}$ is the EGT precoding vector.

polarizations will experience different fading. By combining a number of independently fading antenna elements the probability that all antenna elements are in a fade is reduced, thus creating a more stable communication link. Diversity performance can be characterized by diversity order, which is the effective number of independently fading antenna elements. The diversity order determines the slope of the symbol error rate curve as a function of SNR in a log-log scale. In the case of IID Rayleigh fading, MRT and MRC gives diversity order N_T and N_R, respectively, while dominant eigenmode transmission gives diversity order $N_T N_R$. Note that even in a scattering channel, closely spaced co-polarized antenna elements will have some fading correlation, making the IID assumption invalid.

So far it has been assumed that the channel is known to the transmitter and receiver. Without any channel knowledge, array gain cannot be obtained but diversity gain is still possible to achieve. A simple Tx diversity technique for two Tx antennas that can achieve full diversity order without any channel knowledge at the transmitter is the Alamouti scheme [6]. In this scheme, a complex symbol s_1 is transmitted on a first antenna and another complex symbol s_2 on a second antenna in a first symbol period. In the next symbol period, $-s_2^*$ is transmitted on the first antenna and s_1^* on the second antenna. Assume that the receiver has a single Rx antenna and arranges two consecutive received signal samples in a vector, $\mathbf{y} = [y_1 \; y_2^*]^T$. If the channel is constant over the two symbol periods, the received signal vector is given by

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2^* \end{bmatrix} = \sqrt{P/2}\tilde{\mathbf{H}}\mathbf{s} + \mathbf{n}, \tag{7.7}$$

where

$$\tilde{\mathbf{H}} = \begin{bmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{bmatrix} \tag{7.8}$$

is the effective channel and

$$\mathbf{s} = \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}. \tag{7.9}$$

Since $\tilde{\mathbf{H}}$ is an orthogonal matrix, the receiver can retrieve the transmitted symbols by simply multiplying by $\tilde{\mathbf{H}}^H$ according to

$$\mathbf{z} = \tilde{\mathbf{H}}^H\mathbf{y} = \sqrt{P/2}\left\|\tilde{\mathbf{H}}\right\|_F^2\mathbf{s} + \tilde{\mathbf{H}}\mathbf{n} \tag{7.10}$$

In this way, the full Tx diversity order of two is obtained without any knowledge of the channel at the transmitter. However, no Tx array gain is achieved due to the absence of channel knowledge at the transmitter. The Alamouti scheme has rate one, i.e., two symbols are conveyed over two symbol periods. It is a special case of orthogonal space–time block codes (OSTBCs) for two antennas. OSTBCs exist also for more than two Tx antennas and provide full diversity order with simple linear processing in the receiver. However, for complex signal constellations, an orthogonal code with rate one exists only for two Tx antennas [43]. In more general terms, Tx diversity can be achieved by mapping the modulation symbols in the space–time or space–frequency domains, the so-called space–time transmit diversity (STTD) and space frequency transmit diversity (SFTD), respectively.

Antenna diversity can be combined with frequency diversity if the channel is frequency selective. Otherwise, a frequency-selective channel can be created artificially from a spatially dispersed channel using the so-called delay diversity. This is accomplished by transmitting delayed copies of the same signal on different antennas, thus creating time dispersion or, equivalently, frequency selectivity. If the antennas have uncorrelated fading, the signal at the receiver will appear as a signal that has passed through a channel having multiple taps with uncorrelated fading, i.e., a frequency-selective channel. Delay diversity is transparent to the UE since it only sees the effective frequency-selective channel. Hence, delay diversity can be implemented without any specification support. A particular type of delay diversity that is suitable for OFDM systems is cyclic delay diversity (CDD) [20]. In CDD a cyclic shift instead of a linear delay is applied on the antennas. This is equivalent to applying a frequency dependent phase shift prior to the OFDM modulation.

In LTE, downlink Tx diversity schemes for up to four antennas are supported in the specifications. For two antennas, Tx diversity is based on the so-called space frequency block coding (SFBC), which is equivalent to Alamouti coding in the frequency domain. With four Tx antennas, SFBC is used in combination with the so-called frequency-switched transmit diversity (FSTD) [13]. There is also a downlink transmission mode that combines precoded multilayer transmission with a CDD scheme called large-delay CDD. Uplink Tx diversity using two antennas was introduced for the control channel in LTE Release 10 and uses so-called spatial orthogonal-resource transmit diversity (SORTD) [13]. With SORTD, a signal is transmitted on the different antennas using orthogonal resources in frequency, time, and/or code domain. In NR, however, Tx diversity is currently not explicitly supported and one has to rely on specification-transparent methods.

7.2.2 SPATIAL MULTIPLEXING

As described in the previous section, multiple antennas at the transmitter and receiver can give array gain by precoding and Rx antenna combining. This will increase the SNR which in turn will increase the data rate. This is efficient when the data rate is power limited rather than bandwidth limited. From basic information theory, the achievable data rate grows approximately linearly with the SNR when the SNR is low [46]. However, at high SNR, the achievable rate starts to saturate, since it grows only logarithmically with SNR. In this regime, it would be a more efficient use of the available bandwidth if one could "split" the SNR over several weaker links that could communicate in parallel. Indeed, this is possible by utilizing multiple antennas at the transmitter and receiver and it is the basic principle of spatial multiplexing. To realize this, perform an SVD of a channel matrix $\mathbf{H}$ according to

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H. \tag{7.11}$$

If the transmitter applies a precoder matrix $\mathbf{V}$ and the receiver multiplies the received signal vector by $\mathbf{U}^H$ the effective channel matrix $\tilde{\mathbf{H}}$ becomes

$$\tilde{\mathbf{H}} = \mathbf{U}^H\mathbf{U}\mathbf{\Sigma}\mathbf{V}^H\mathbf{V} = \mathbf{\Sigma}. \tag{7.12}$$

Since $\mathbf{\Sigma}$ is a diagonal matrix, the effective channel is composed of multiple parallel subchannels without any crosstalk, where the gain of each subchannel is determined by the corresponding singular value. Independent data streams, or layers, can then be transmitted over the subchannels without any mutual

interference. In this way, the achievable data rate can increase linearly with the number of antennas,[12] thus circumventing the data rate saturation at high SNR. This approach is deceptively simple but involves several issues that need to be solved in a practical system. For example, it requires that the channel matrix is known to both the transmitter and the receiver. In practice, this cannot be known exactly and there will be some interference between the transmitted data layers. More advanced receivers can then be used to suppress this interference; see Section 7.2.2.3.

The performance of spatial multiplexing depends to a large extent on the channel properties. If the channel has a high correlation, some singular values will be small, leading to weak subchannels that will not give any significant contribution to the overall data rate. The number of non-zero singular values of the channel matrix is called the channel rank.[13] In a practical system it is important to dynamically adapt the number of used subchannels to the current channel and data traffic conditions in order to optimize the overall data rate, so-called rank adaptation. The number of used subchannels is often referred to as the transmission rank. The transmission rank can depend on the channel rank, but it can also depend on other parameters such as the SINR. To make use of many subchannels, i.e., a high transmission rank, a high channel rank may not be sufficient. A high SINR is usually also required. To benefit from spatial multiplexing, a high "basic" SINR is needed, since the transmission power has to be shared between the transmitted layers and in practice there will be some interference between the layers.

Precoding and spatial multiplexing can be combined. One example was given above where premultiplication of $\mathbf{V}$ orthogonalizes the channel on the Tx side which makes the processing at the Rx side easier. If the number of transmitted layers is lower than the number of Tx antenna elements, precoding can also increase the SNR per layer by providing array gain. In order to determine a suitable precoder matrix, CSIT is required. For example, the channel matrix can be estimated from uplink SRSs and the precoder matrix can be determined from an SVD of the estimated channel matrix. However, this requires TDD operation and calibration between uplink and downlink RF branches. For FDD, a common approach is to let the UE estimate the channel based on downlink reference signals and feed back a proposed precoder matrix. To save feedback overhead, a limited number of predefined precoder matrices can be collected in a so-called codebook. The UE then needs only to signal an index to the preferred matrix in the codebook, rather than the complexed-valued coefficients of the precoder matrix itself. The codebook can contain precoder matrices for different numbers of transmission ranks. The codebook needs to be known by both the BS and the UE and thus requires specification support. This scheme is referred to as codebook-based precoding and has been adopted in both the LTE and the NR specifications. More details on how these codebooks are constructed are given in Section 7.2.6.

Spatial multiplexing of several layers to/from a single UE is often referred to as SU-MIMO. The layers may also be multiplexed to/from different UEs. This is called MU-MIMO or space division multiple access (SDMA). SU-MIMO and MU-MIMO can be combined so that the spatially multiplexed UEs can have multiple layers each. Residual interference between the transmitted layers can be suppressed by the receiver as described in Section 7.2.2.3. However, this is more difficult in downlink MU-MIMO, since Rx antennas from different UEs cannot be processed coherently and the number of antenna elements per UE might not be sufficient to suppress all interference. Downlink MU-MIMO

[12] More precisely, under certain conditions the achievable rate grows as $\min\{N_R, N_T\}$.

[13] In practice, no singular values will be exactly zero so some kind of threshold could be used.

transmission may therefore require some interference suppression at the Tx side. In a cellular system the channel conditions for different UEs will vary rapidly with time. To optimize system performance it is therefore important to dynamically switch between SU- and MU-MIMO operation depending on channel conditions and traffic load.

7.2.2.1 SU-MIMO Precoding

The previous discussion of precoding concerned maximizing the SNR for a single layer. As alluded to previously, precoding and spatial multiplexing can be combined. We can then find the precoder that maximizes the sum rate of all layers. Assuming that the channel is known to both the transmitter and the receiver, the capacity-optimal precoder under a sum-power constraint is given by [32]

$$\mathbf{W}_T = \mathbf{V}\mathbf{P}^{1/2} \tag{7.13}$$

where $\mathbf{V}$ is obtained from the SVD of $\mathbf{H}$ according to (7.11) and $\mathbf{P} = \text{diag}\{p_1, \ldots, p_{N_T}\}$ is a diagonal matrix containing the power allocated to each layer. The optimal power allocation is obtained from the well-known waterfilling algorithm; see [46]. This algorithm will allocate high power to strong subchannels. The capacity-optimal transmission scheme under a per-antenna power constraint is a more difficult problem and one may need to resort to numerical techniques to find the optimal solution [41]. A closed-form solution for the general problem seems to be unknown, but for the case when the channel matrix has full column rank and the input covariance matrix has full rank, the optimal input covariance is given by [48]

$$\mathbf{R}_{\text{opt}} = \mathbf{I} + \text{diag}\left[\left(\mathbf{H}^H\mathbf{H}\right)^{-1}\right] - \left(\mathbf{H}^H\mathbf{H}\right)^{-1} \tag{7.14}$$

where $\mathbf{I}$ denotes the identity matrix. The optimal precoder matrix is then obtained from

$$\mathbf{W}_T = \mathbf{E}\mathbf{\Lambda}^{1/2} \tag{7.15}$$

where $\mathbf{E}$ and $\mathbf{\Lambda}$ are obtained from the eigendecomposition $\mathbf{R}_{\text{opt}} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^H$.

Obtaining knowledge of the instantaneous channel matrix at the transmitter may be difficult in a practical system, e.g., if the channel is varying rapidly due to UE movement. A precoder design can then be based on channel statistics, e.g., the Tx channel covariance matrix, $\mathbf{R}_T = \text{E}[\mathbf{H}^H\mathbf{H}]$, where $\text{E}[\cdot]$ denotes expectation, instead of the instantaneous channel. Assuming that the Tx covariance matrix is known to the transmitter, the optimal[14] precoding matrix under a sum-power constraint is given by [23,47]

$$\mathbf{W}_T = \mathbf{E}\mathbf{P}^{1/2} \tag{7.16}$$

where $\mathbf{E}$ is now obtained from the eigendecomposition of $\mathbf{R}_T = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^H$ and $\mathbf{P}$ is the diagonal matrix that allocates the power over the eigenvectors of $\mathbf{R}_T$. An iterative method to find $\mathbf{P}$ has been proposed in [47]. An approximate solution is to apply "statistical waterfilling", i.e., to use the eigenvalues of $\text{E}[\mathbf{H}^H\mathbf{H}]$ instead of $\mathbf{H}^H\mathbf{H}$ as is used in the conventional waterfilling algorithm.

[14] Optimal with respect to ergodic capacity.

7.2.2.2 MU-MIMO Precoding

In this section, we discuss precoding for MU-MIMO transmission. There are several aspects that make MU-MIMO different from SU-MIMO [32]:

- SU-MIMO performance can be characterized by a link capacity while MU-MIMO performance is characterized in terms of a capacity region, i.e., the set of simultaneously achievable rates for all UEs.
- In SU-MIMO, only the sum rate of all layers is of interest since all layers are transmitted to the same user. In MU-MIMO, the layers are transmitted to different UEs and fairness between UEs needs also to be taken into account.
- In SU-MIMO, the transmission loss for all Tx–Rx antenna pairs is usually similar, while there can be a large difference in transmission loss to different UEs in MU-MIMO.
- In SU-MIMO, the Rx antenna elements can be combined coherently to optimize performance, e.g., by suppressing interlayer interference. Rx antennas in different UEs cannot be combined coherently in MU-MIMO.
- MU-MIMO requires more accurate channel knowledge at the transmitter than SU-MIMO. One reason for this is that interlayer interference needs to be suppressed at the transmitter if the number of transmitted layers is larger than the number of Rx antennas in the UEs and interference suppression requires accurate channel knowledge.

For simplicity, we assume that all UEs have a single antenna element each. Furthermore, we assume that the channel is known to both the transmitter and the receivers. The signal received by the kth out of K co-scheduled UEs served by a BS with N_T Tx antenna elements can be modeled by

$$y_k = \mathbf{h}_k \sum_{i=1}^{K} \mathbf{w}_i s_i + n_k \tag{7.17}$$

where $\mathbf{h}_k$ is the $1 \times N_T$ channel vector to UE k, $\mathbf{w}_i$ is the $N_T \times 1$ precoding vector to UE i, s_i the signal transmitted to UE i, and n_k is additive receiver noise with power N_0. We assume that $\|\mathbf{w}_i\|_F^2 = P_i$ where P_i is the power allocated to UE i and $\mathrm{E}[|s_i|^2] = 1$. The SINR of UE k is thus given by

$$\mathrm{SINR}_k = \frac{|\mathbf{h}_k \mathbf{w}_k|^2}{\sum_{i \neq k} |\mathbf{h}_k \mathbf{w}_i|^2 + N_0}. \tag{7.18}$$

Finding the optimal linear precoder may be posed as maximizing some utility function[15] $f(\mathrm{SINR}_1, \ldots, \mathrm{SINR}_K)$ subject to a constraint on the total transmitted power according to $\sum_{i=1}^{K} P_i \leq P$. In general, this is a very difficult problem. However, the solution has a simple general structure according to [9]:

$$\mathbf{w}_{k,opt} = c\sqrt{P_k} \left(\mathbf{I} + \sum_{i=1}^{K} \frac{\lambda_i}{N_0} \mathbf{h}_i^H \mathbf{h}_i \right)^{-1} \mathbf{h}_k^H = c\sqrt{P_k} \left(\mathbf{I} + \frac{1}{N_0} \mathbf{H}^H \mathbf{\Lambda} \mathbf{H} \right)^{-1} \mathbf{h}_k^H \tag{7.19}$$

[15] One example of a utility function is the sum rate $f(\mathrm{SINR}_1, \ldots, \mathrm{SINR}_K) = \sum_{i=1}^{K} \log_2(1 + \mathrm{SINR}_i)$.

for some positive parameters $\lambda_1, \ldots, \lambda_K$ such that $\sum_{i=1}^{K} \lambda_i = P$. Here, $\mathbf{w}_{k,opt}$ is the optimal precoding vector for UE k, c is a normalization that makes $\|\mathbf{w}_{k,opt}\|^2 = P_k$, $\mathbf{H} = [\mathbf{h}_1^T \ \cdots \ \mathbf{h}_K^T]^T$, and $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_K\}$. In general it is difficult to find the optimal λ_i in closed form. However, several well-known precoders correspond to specific choices of λ_i such as the regularized ZF [33], minimum mean square error (MMSE), transmit Wiener filter [24], and signal-to-leakage-and-interference ratio (SLNR) [39] precoders. For example, letting $\lambda_i = P/K$ for all UEs leads to the MMSE solution

$$\mathbf{w}_k = c\sqrt{P_k}\left(\mathbf{I} + \frac{P}{KN_0}\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{h}_k^H. \tag{7.20}$$

Other well-known precoders can also be obtained from (7.19) asymptotically for low and high SNR. At low SNR ($N_0 \to \infty$), (7.19) reduces to

$$\mathbf{w}_k = c\sqrt{P_k}\mathbf{h}_k^H, \tag{7.21}$$

i.e., the MRT precoder in (7.4). The MRT MU-MIMO precoder "beam-forms" a layer to its intended UE, while ignoring interference to co-schedules UEs. At high SNR ($N_0 \to 0$), we obtain the ZF precoder

$$\mathbf{w}_k = c\sqrt{P_k}\mathbf{h}_k^{\dagger} \tag{7.22}$$

where $\mathbf{h}_k^{\dagger}$ denotes the kth column of $\mathbf{H}^{\dagger}$, and $\mathbf{H}^{\dagger} = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H)^{-1}$ is the pseudo-inverse of $\mathbf{H}$. The ZF precoder beam-forms a layer in the direction to the desired UE while placing nulls in the directions to co-schedules UEs.[16] Since this precoder places nulls at co-scheduled UEs, the UEs will receive no inter-user interference. However, if the channel vectors for two different UEs are close to parallel, there will be a gain reduction that leads to a loss in SNR. The regularization of the inverse in (7.19) and (7.20) mitigates this problem and provides a balance between gain reduction to the desired UE and interference suppression, which leads to good performance over a wide SNR range.

Linear precoding techniques suffer from performance degradation when the channels to co-scheduled UEs are highly correlated. This can be improved by using nonlinear precoding techniques. The optimal approach is the so-called dirty paper coding (DPC) which achieves the maximum sum rate [51]. The idea with DPC is to precancel interference at the transmitter by using perfect knowledge about the channel and transmitted signals. It rests on the fundamental result in [12] that interference does not decrease the channel capacity if the transmitter knows the interference at the receiver, even if the receiver has no knowledge regarding the interference.

DPC-based precoding can be used in MU-MIMO to eliminate interference between UEs through coding and interference presubtraction [10]. To realize the viability of such an approach, let $\mathbf{H} = \mathbf{R}\mathbf{Q}$ be the QR decomposition of the channel matrix. Here, $\mathbf{R}$ is a $K \times K$ lower triangular matrix and $\mathbf{Q}$ is a $K \times N_T$ unitary matrix, i.e. $\mathbf{Q}\mathbf{Q}^H = \mathbf{I}$. By using the precoder

$$\mathbf{W}_T = \mathbf{Q}^H \tag{7.23}$$

[16] This is a LoS description, but it holds also for multipath channels if "direction" is interpreted in a wider sense.

the symbols received at the UEs are given by

$$\mathbf{y} = \sqrt{P}\mathbf{R}\mathbf{s} + \mathbf{n}. \tag{7.24}$$

Since $\mathbf{R}$ is lower triangular, the received symbol for the kth UE is, ignoring the additive noise term,

$$y_k = \sum_{i=1}^{k} r_{i,k} s_i, \; k = 1, \ldots, K, \tag{7.25}$$

where $r_{i,k}$ denotes the (i,k)th element of $\mathbf{R}$. Hence, the first UE receives no interference, the second UE receives interference only from the first UE, and so on. Interference presubtraction can then be achieved by replacing the transmitted symbols $s_k, \; k = 1, \ldots, K$, by

$$s'_k = s_k - \frac{1}{r_{k,k}} \sum_{i=1}^{k-1} r_{k,i} s'_i, \tag{7.26}$$

so that the UEs receive the symbols

$$y_k = \; r_{k,k} s_k, \; k = 1, \ldots, K. \tag{7.27}$$

Hence, the interference presubtraction has completely removed the interference between UEs. This scheme can be seen as the Tx equivalent of the successive interference cancellation (SIC) receiver described in Section 7.2.2.3. An advantage with performing the interference subtraction on the Tx side is that it does not suffer from error propagation since the transmitted signals are known.

DPC precoding is, however, complex and may be difficult to implement in practice. Numerous suboptimal, nonlinear precoding techniques with lower complexity have therefore been developed, e.g. the Tomlinson–Harashima method [14], and Vector Perturbation [18] precoding.

The MU-MIMO precoders described so far have been developed under the assumptions of a single-cell system with perfect channel knowledge at the transmitting BS. In a practical system the performance of such precoders is impaired by channel estimation errors and intercell interference. To mitigate this, the MMSE precoder can be generalized to take channel estimation errors and intercell interference into account [25,27,28]. Such a multicell MMSE precoder has roughly the general structure (see the cited references for details that have been omitted here for the sake of brevity)

$$\mathbf{w}_k \sim \left(\mathbf{I} + \alpha \hat{\mathbf{H}}_{\text{intra}}^H \hat{\mathbf{H}}_{\text{intra}} + \beta \hat{\mathbf{H}}_{\text{inter}}^H \hat{\mathbf{H}}_{\text{inter}} + \gamma \mathbf{C} \right)^{-1} \hat{\mathbf{h}}_k^H \tag{7.28}$$

where $\hat{\mathbf{H}}$ denotes estimated channels and $\mathbf{C}$ a channel estimation error covariance matrix. The different terms inside the inverse account for intracell interference, intercell interference, and channel estimation errors, respectively. The multicell MMSE precoder suppresses interference to co-scheduled UEs within the served cell as well as the interference to UEs in other cells, while taking the channel estimation quality into account.

7.2.2.3 MIMO Receivers

This section gives a brief overview of different MIMO receivers that can be used to suppress interference. They are presented in the context of interlayer interference suppression in spatial multiplexing, but the same principles apply for canceling other types of interference, e.g., intercell interference.

The optimal approach to MIMO receiver design is the maximum likelihood (ML) principle. The ML receiver searches over all possible transmitted signal vectors to find the most likely one. This is a nonlinear receiver which usually is too complex to implement. A simpler nonlinear receiver is the so-called SIC receiver [49]. The idea with SIC is to successively demodulate and decode the different layers, and to reencode and subtract their contributions to the received signal, layer by layer. Performance may be impaired by error propagation, which occurs if a layer is not decoded correctly. This can be mitigated by ordering the layers so that the layer with highest SINR is decoded in each stage, so-called ordered SIC.

To reduce complexity further, linear receivers can be used. A linear receiver applies a linear filter to separate the transmitted layers and then decodes each layer independently. The output of a linear receiver can be expressed as

$$\mathbf{z} = \mathbf{W}_R \mathbf{y} \tag{7.29}$$

where $\mathbf{z}$ is the $N_T \times 1$ vector output of the receiver, $\mathbf{W}_R$ an $N_T \times N_R$ Rx weight matrix, and $\mathbf{y}$ is the $N_R \times 1$ received signal vector before the receiver. The received signal vector can be modeled as

$$\mathbf{y} = \sqrt{P}\mathbf{H}\mathbf{s} + \mathbf{n} \tag{7.30}$$

where $\mathbf{H}$ is the $N_R \times N_T$ channel matrix,[17] $\mathbf{s}$ is the $N_T \times 1$ transmitted signal vector, and $\mathbf{n}$ is a noise vector, which is assumed to be spatially white with covariance matrix $N_0\mathbf{I}$.

A ZF receiver removes interlayer interference by inverting the channel according to

$$\mathbf{W}_R = \frac{1}{\sqrt{P}}\mathbf{H}^{\dagger} \tag{7.31}$$

where $\mathbf{H}^{\dagger} = (\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H$ is the pseudo-inverse of $\mathbf{H}$. The output of the ZF receiver is given by

$$\mathbf{z} = \mathbf{s} + \frac{1}{\sqrt{P}}\mathbf{H}^{\dagger}\mathbf{n}. \tag{7.32}$$

Hence, the ZF receiver eliminates the interference completely and decouples the matrix channel into N_T parallel scalar channels. However, it also increases the noise level at the output of the receiver since the noise vector $\mathbf{n}$ is premultiplied by $\mathbf{H}^{\dagger}$. If $\mathbf{H}$ is close to singular, $\|\mathbf{H}^{\dagger}\|_F$ will be large and the noise increase will be large. Furthermore, $\mathbf{W}_R$ makes the noise spatially colored at the output of the receiver.

[17] We ignore any potential precoding at the transmitter. In the case of precoding, the precoding matrix could be absorbed in $\mathbf{H}$ to represent an effective channel matrix. N_T is then the number of transmitted layers, which could be lower than the number of Tx antenna elements.

The SINR of the kth layer on the output of the ZF receiver is given by

$$\text{SINR}_k = \frac{P}{N_0} \frac{1}{\left[\left(\mathbf{H}^H\mathbf{H}\right)^{-1}\right]_{k,k}} \tag{7.33}$$

where $[\cdot]_{k,k}$ denotes the kth diagonal element of a matrix. It can be shown that the diversity order and array gain for each layer is proportional to $N_R - N_T + 1$ [32].

The noise enhancement of the ZF receiver can be mitigated by suppressing the interference down to the noise level instead of canceling it completely. This is accomplished by the MMSE receiver which is derived by minimizing the squared error between the received and transmitted signal vector. The solution is

$$\mathbf{W}_R = \frac{1}{\sqrt{P}} \left(\mathbf{H}^H\mathbf{H} + \frac{N_0}{P}\mathbf{I}\right)^{-1} \mathbf{H}^H. \tag{7.34}$$

The MMSE receiver balances noise enhancement and interference suppression to minimize the total power of interference and noise. From (7.34) it can be seen that the MMSE solution converges to ZF for high SNR and to MRC for low SNR. The SINR of the kth layer on the output of the MMSE receiver is given by [32]

$$\text{SINR}_k = \frac{1}{\left[\left(\frac{P}{N_0}\mathbf{H}^H\mathbf{H} + \mathbf{I}\right)^{-1}\right]_{k,k}} - 1. \tag{7.35}$$

7.2.3 ANTENNA ARRAY ARCHITECTURES

As the name suggests, multiantenna techniques require multiple antennas. Although not necessary, these are often collocated in a single enclosing structure. Multiantenna techniques can also be applied across several such enclosures of antenna elements, placed in different locations. Herein, we refer to a collection of collocated antenna elements as an antenna array. This section gives a high-level description of some different architectures for antenna arrays.

NR is expected to operate in a large span of carrier frequencies: from below 1 GHz up to 100 GHz. As with many other technology components also the antenna system designs will be different in different parts of this vast frequency span. This is partly due to building practices and hardware implementation issues and partly due to system level and propagation aspects. Without going into implementation details, this section discusses different antenna array architectures that are viable candidates in different parts of this frequency range. Although a 3GPP technical specification rarely dictates a particular hardware implementation or architecture, it has in several cases been developed to be suited for relevant implementations. Therefore, some parts of the specifications are suitable for certain antenna array architectures, while others are not.

7.2.3.1 Digital Arrays

In a digital array architecture each antenna element is equipped with its own RF chain and data converters (ADC and DAC). Fig. 7.2 shows a schematic illustration of the transmitting part of a digital

array architecture. The receiving part would look the same if the DACs are replaced with ADCs. The antenna array is depicted as a 1-D linear array but it could have any topology, e.g., a 2-D planar array. It is common practice that each antenna element position is populated with two radiating elements having orthogonal polarization,[18] each polarization having its own RF chain and DAC. However, for ease of exposition, this is omitted here. The illustration should be seen as functional rather than an implementation description. Digital arrays designed for OFDM systems also have an FFT/IFFT for each antenna element, enabling frequency-selective precoding. In principle, different precoding weights can thus be applied for each subcarrier. In practice, however, this also requires CSI for every subcarrier, which may not always be available due to, e.g., constraints on the frequency granularity in the CSI feedback in an FDD system. Therefore, frequency-selective precoding is often used with a subband granularity, where a subband contains several consecutive subcarriers. Besides being limited by CSI, the granularity of frequency-selective precoding can also be limited by the signal processing capacity, since calculating precoding weights per subcarrier can be computationally demanding.

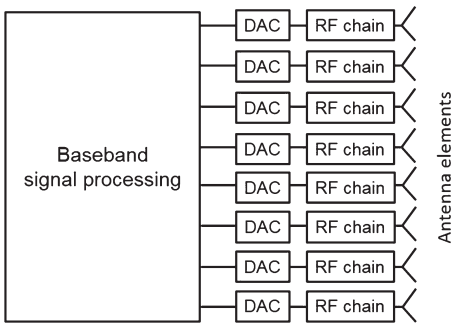


FIGURE 7.2

Digital array architecture. Schematic illustration of the digital array architecture.

Advances in active array antenna technology have made it possible to produce digital arrays with a large number of elements. Having an RF chain and data converter for each antenna element provides the highest performance and flexibility. Multiantenna techniques such as spatial multiplexing and interference suppression can reach its full potential when used in a fully digital array. However, a fully digital array can also be expensive in terms of cost and power consumption. These aspects are particularly pronounced at millimeter-wave frequencies, since the number of antenna elements is expected to be large in order to populate a sufficiently large physical antenna area that can achieve a required link budget. Furthermore, the large bandwidth foreseen to be used at these frequencies requires the data converters to operate at high sampling rates, leading to high power consumption and heat generation. A large bandwidth coupled with many digitized antenna elements is also challenging from a shear data shuffling perspective, putting high demands on data interfaces between the antenna array and signal processing units. This also leads to high demands on signal processing capacity. Therefore, a fully digital array is currently a likely implementation only in the low frequency bands. In millimeter-wave bands, analog and hybrid array architectures will be prevalent, at least in the near future. These architectures are described in the sequel.

[18] Physically, it may be a single element with different excitation points.

7.2.3.2 Analog Arrays

A schematic illustration of the transmitting part of an analog array architecture is shown in Fig. 7.3. With an analog array, analog beam-forming can be performed by applying a linear phase progression over the array by means of phase shifters[19] in order to steer a beam in the desired direction. If the array has some form of gain control per antenna element, amplitude tapering can be applied to reduce the sidelobes in order to mitigate interference to/from other UEs.

With an analog array, beam-forming is usually limited to wideband beam-forming, i.e., the same beam-forming weights are used over the entire bandwidth. With wideband beam-forming it is therefore not possible to adapt to the frequency selectivity of the channel. However, the spatial characteristics of the channel, such as the main direction of energy is typically not frequency dependent. Analog beam-forming is therefore suited for scenarios where the channel energy has a dominant direction, e.g., scenarios with LoS, a strong specular reflection or a dominating cluster with low angular spread.

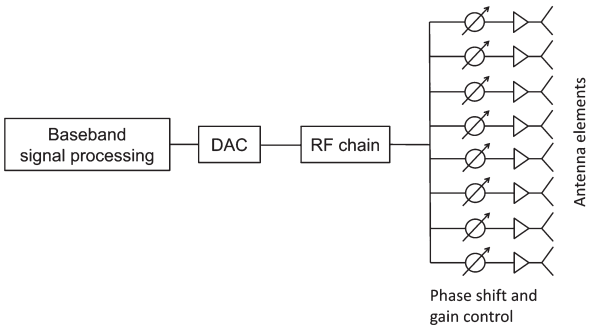


FIGURE 7.3

Analog array architecture. Schematic illustration of the analog array architecture.

A digital array has higher performance potential than an analog array since it provides more degrees of freedom in the spatial signal processing. However, at millimeter-wave frequencies many degrees of freedom may not be crucial to the system performance. It is likely that millimeter-wave systems will be deployed in small cells, at least initially. Since small cells have a high probability of LoS, the benefits with frequency-selective digital precoding over analog beam-forming becomes small since the angle and delay spread is small. In a LoS channel with no angle or delay spread, the optimal precoder is simply a wideband beam-former if interference is ignored. This can be implemented by an analog array. Efficient interference nulling is difficult to implement with analog arrays, but amplitude tapering can be performed in the analog beam-forming to reduce sidelobe levels and hence interference.

An important limitation of the analog array architecture from a system design point of view is that an analog beam-former can only transmit or receive in one direction at a time. Indeed, this has had a profound impact on the development of the NR specifications. In order to provide support for analog beam-forming of both data and control channels, a set of procedures called *beam management* has been developed. This is described in Section 7.3.4. Also the initial access procedures have been designed to support analog beam-forming.

[19] For systems with large relative bandwidth, phase shifts should be replaced by true time delays.

7.2.3.3 Hybrid Arrays

The fully digital and analog array architectures presented in previous sections represent two extreme cases, each having their advantages and disadvantages with regard to performance, cost, and complexity. In a hybrid array architecture, more flexibility in these trade-offs is provided by combining the digital and analog array architectures. Two examples of a hybrid array architecture are shown in Fig. 7.4, referred to as fully and partially connected, respectively. In the fully connected architecture, multiple (in this case two) arrays of phase shifters and gain controllers have the same antenna elements. In this example, two beams with full beam-forming gain can be generated independently of each other, since each analog chain has its own phase shifters and gain controllers and they are connected to all antenna elements. These two beams can then be combined digitally in the baseband signal processing, for example to perform spatial multiplexing or to increase the array gain in a multipath environment.

The partially connected architecture is a less complex hybrid architecture in which each analog chain is connected to a subarray of antenna elements. A drawback with this architecture compared to the fully connected architecture is that each analog beam-former cannot achieve the full beam-forming gain, only the gain provided by an individual subarray. However, full array gain may be achieved by digital beam-forming over the analog beams created by each subarray, so-called hybrid beam-forming. The analog and digital beam-former should then point in the same direction, otherwise grating lobes will appear and the gain is reduced. Hybrid beam-forming may also require some calibration between the analog subarrays.

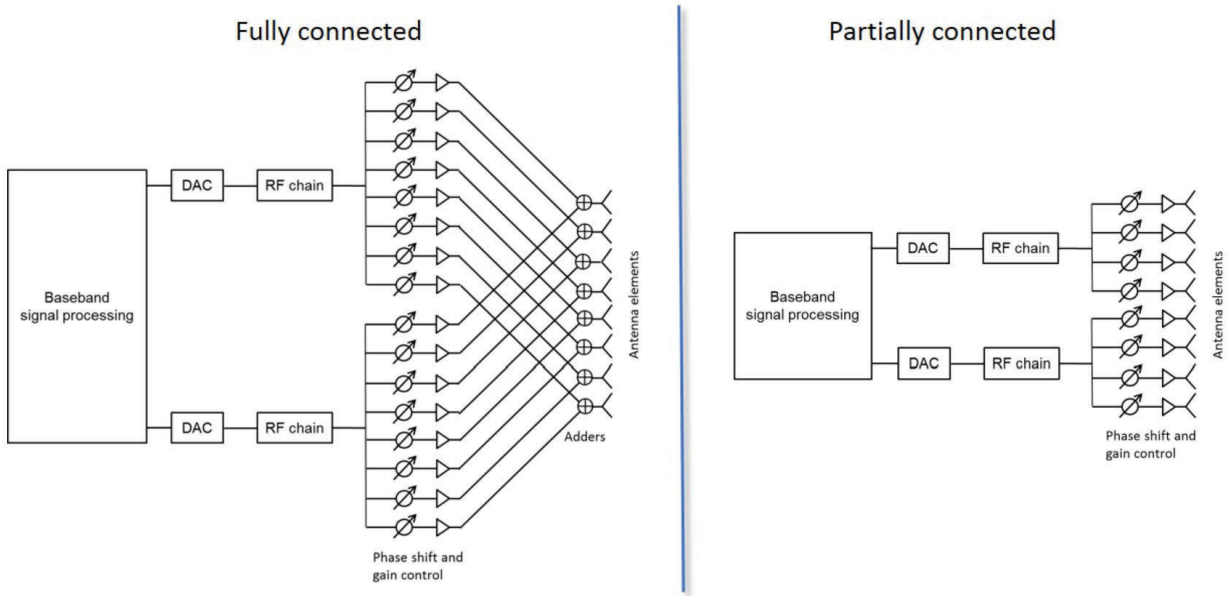


FIGURE 7.4

Hybrid array architecture. Schematic illustration of hybrid array architectures.

Drawbacks with the fully connected architecture compared to the partially connected architecture are higher complexity and losses in the adders and dividers (or combiners in an Rx array). Clearly, the fully connected array has more components due to the adders and the replication of phase shifters. The

adders also introduce RF losses that scale linearly with the number of added signals. Furthermore, the losses in the dividers between the RF chains and phase shifters also scale linearly with the number of phase shifters, which is larger in the fully connected architecture [16].

With a hybrid array architecture, the channel at antenna element level cannot be observed due to the analog beam-forming. A challenge is then how to co-design the analog and digital beam-formers. For the highest performance they should be designed jointly, but this may be too complex. A suboptimal but simpler approach is to design the beam-formers independently; see Section 7.3.4.3.

A partially connected hybrid array architecture with rectangular subarrays that has been evaluated extensively in the development of the NR specifications is the so-called panel array [1,3]. In 3GPP, the subarrays are referred to as panels and each panel is a uniform planar array (UPA) with single- or dual-polarized elements. An example of panel array with 2×2 panels, each with 4×4 dual-polarized antenna elements is illustrated in Fig. 7.5. A possible implementation is that analog beam-forming is performed per polarization in each panel and that each panel has two RF chains, one per beam/polarization. The panels can be used in different ways. For example, they can be used independently to serve different UEs or be combined coherently to serve a single UE.

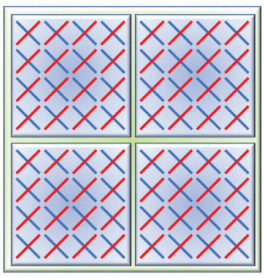


FIGURE 7.5

Panel array. An example a panel array with 2×2 panels, each with 4×4 dual-polarized antenna elements.

7.2.3.4 A Millimeter-Wave Antenna Array System Prototype

An example of a compact millimeter-wave active antenna array system is shown in Fig. 7.6. The upper left picture shows the top view of an active array antenna module for 28 GHz carrier frequency designed by Ericsson and IBM Watson Research Center [35,40,15]. One module has 64 dual-polarized antenna elements[20] and can generate one beam per polarization using analog beam-forming. Each antenna element and polarization has its own front-end radio with phase shifter and gain control, i.e., 128 front-end radio chains in one module. The size of one antenna array is 70 × 70 mm. The lower left picture shows the bottom view of a module with four radio frequency integrated circuits (RFICs). Each RFIC has 32 Tx and Rx branches containing mixer, phase shifter, attenuator, PA, low-noise amplifier (LNA), and TDD switches. The right picture shows a radio unit prototype designed by Ericsson that consists of two such modules.

[20] As can be seen in the picture, a module actually has 100 elements but the edge elements are just dummy elements to reduce border effects.

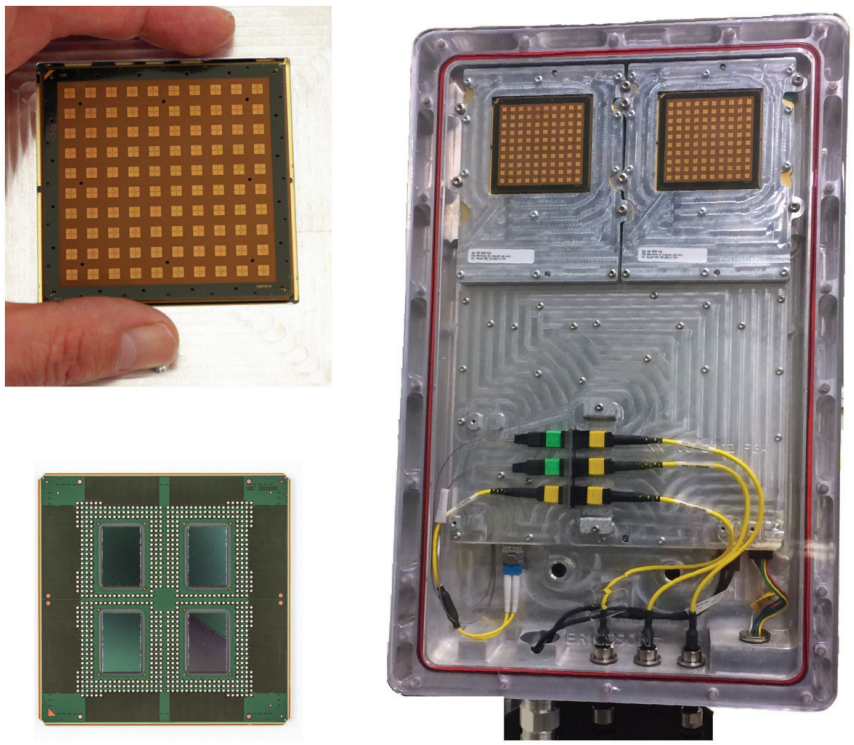


FIGURE 7.6

28 GHz active antenna array. Top left: Front view of one module with 8×8 dual-polarized active antennas. Low left: Back view of one module with four RFICs. Right: A prototype radio unit with two modules.

A high-level architecture of the Tx branches of a radio unit and a connected baseband unit is shown in Fig. 7.7. The receiver architecture is similar but with ADCs instead of DACs and FFTs instead of IFFTs. A radio unit is capable of generating four beams independently, one beam per module and polarization, each having its own ADC/DAC and digital radio. Frequency domain data are transmitted between the radio and baseband unit over a common public radio interface (CPRI). The bandwidth of the prototype is 800 MHz and it is capable of delivering four data streams with a total of 15 Gbps peak data rate. A beam from a module has 23 dBi gain and a half-power beam width of 12° in both azimuth and elevation. Measured azimuth beam patterns for five different beams from one module are shown in Fig. 7.8.

An Ericsson 5G Testbed system including the 28 GHz radio unit has been used in a 5G trial network in Pyeongchang, Republic of Korea, in a cooperation between KT Corporation, Ericsson and several other technology partners [34].

7.2.4 UE ANTENNAS

Antennas of hand-held devices are usually designed to have as close to omni-directional coverage as possible since the incident waves may come from any direction. At traditional cellular frequencies, the size of a typical hand-held device is of the same order as the carrier wavelength. At millimeter-wave

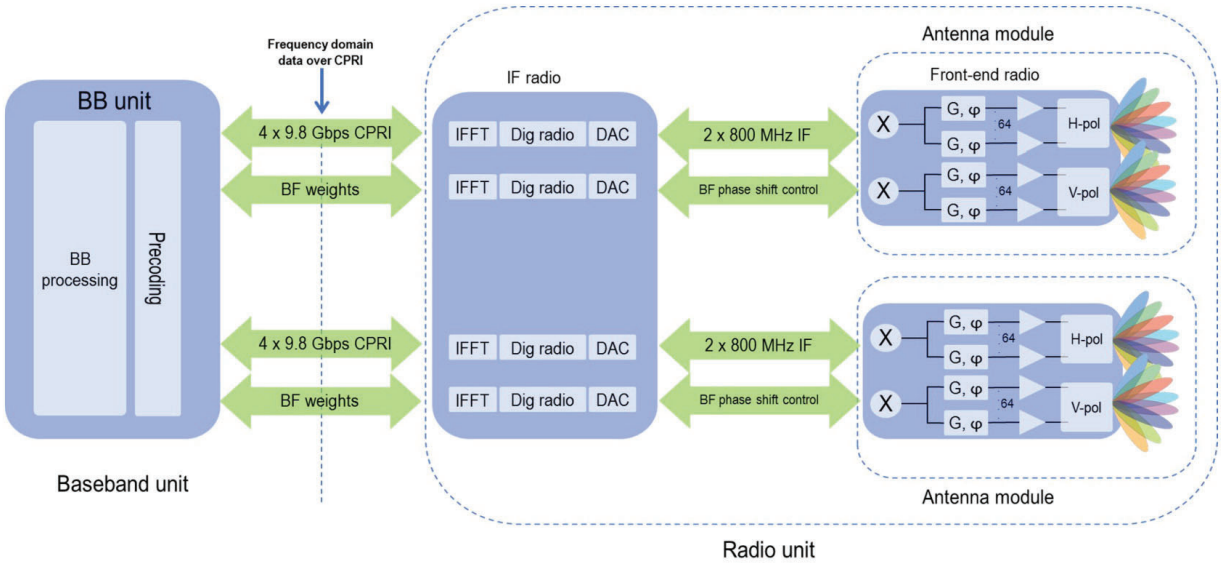


FIGURE 7.7

Architecture. High-level architecture of radio and baseband unit.

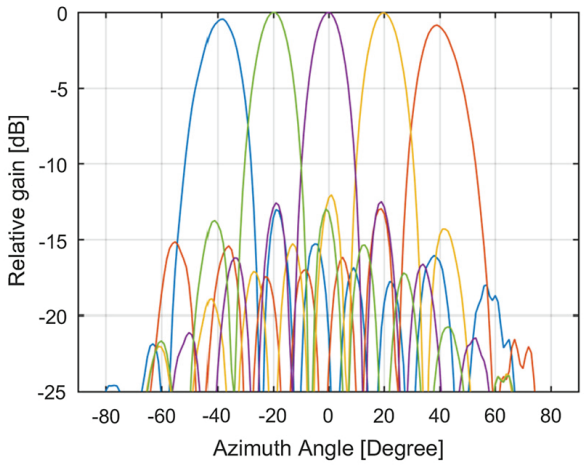


FIGURE 7.8

Beam patterns. Measured azimuth beam patterns for five different beams from one module.

frequencies, however, a hand-held device is large compared to the wavelength. This makes it more difficult to design omni-directional antennas since the device chassis acts as a larger ground plane. On the other hand, since the physical size of an antenna having a certain gain decreases with increasing frequency, it is possible accommodate millimeter-wave antenna arrays to the limited form factor of a hand-held device. One approach to achieve omni-directional coverage at high frequencies is to have several directive antennas that cover different angular sectors. The antennas can be discrete antennas

with a fixed beam pattern or antenna arrays performing dynamic beam-forming. If instantaneous omni-directional coverage is needed each antenna needs its own transceiver, otherwise a single transceiver can be switched to the currently best antenna.

Having antennas at different locations on the device can mitigate the effects of blockage caused by the user putting the hand or finger over an antenna. Furthermore, having different angular coverages of the antennas can reduce the impact of other nearby obstacles such as the user's body, other people, cars, etc. In such cases it is important to be able to quickly switch to another direction to find an alternative propagation path, e.g., a strong reflection [7]. Having different angular coverage of the antennas can also be beneficial for spatial multiplexing and diversity.

If the device is equipped with antenna arrays, dynamic beam-forming can be used to compensate for movements and rotation of the device. Due to stringent requirements on cost and power consumption in hand-held devices, implementation by analog arrays is most likely. Providing support for the UE to dynamically track the best signal direction with analog beam-forming is one of the major new challenges in NR within the multiantenna area. This support is given by the beam management framework, described in more detail in Section 7.3.4.

The feasibility of integrating antenna arrays into mobile phones at millimeter-wave frequencies has been demonstrated with several prototypes. For example, [19] shows a 28 GHz prototype where two 16-element arrays have been integrated on the top and bottom of a cellular phone. Another example of a 28 GHz prototype is shown in [52] where two linear antenna arrays with eight elements each have been integrated on opposite edges of a mobile phone mock-up with metallic back casing. Fig. 7.9 shows a schematic illustration of the antenna array placement on the phone in [19] and [52], respectively.

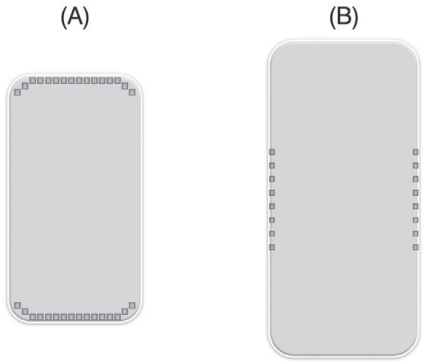


FIGURE 7.9

UE antenna arrays. Prototype millimeter-wave antenna array placement on a phone: (A) in [19], (B) in [52].

A major challenge in UE antenna design is the large number of frequency bands that needs to be supported by the UE. With the exploitation of new bands in millimeter-wave spectrum this number will increase even further. Another challenge is that regulatory requirements on RF electromagnetic field (EMF) exposure of humans are different at millimeter-wave frequencies compared to those in traditional cellular frequency bands. Although the regulations vary between different parts of the world it is common that at frequencies above a transition frequency, typically 6 or 10 GHz, the restriction metric changes from specific absorption rate (SAR) to free-space power density. This change in met-

ric implies that the maximum permissible transmitted power for a UE may be significantly lower at millimeter-wave bands than at frequencies used for current cellular technologies [11].

7.2.5 ANTENNA PORTS AND QCL

Although multiantenna features are treated extensively in 3GPP specifications, the antennas are rarely described as hardware components. Features and procedures are instead referring to so-called *antenna ports*. An antenna port is in the 3GPP specifications an abstract concept that is a logical entity rather than a physical antenna. To ease the understanding of the multiantenna techniques in 3GPP, this section explains the meaning of an antenna port and the related concept of quasicolocation (QCL) and also gives some motivation for their definitions.

In many cases it is important for the receiver to know which assumptions it can make on the channel corresponding to different transmissions. For example, the receiver needs to know which reference signal transmission it can use to estimate the channel in order to decode a transmitted signal. It is also important for the UE to be able to report relevant CSI to the BS which it can use for scheduling and link adaptation purposes. For this purpose, two important concepts were introduced in LTE: antenna port and QCL. An antenna port by definition functions such "that the channel over which a symbol on the antenna port is conveyed can be inferred from the channel over which another symbol on the same antenna port is conveyed" [4]. The receiver can assume that two transmissions correspond to the same radio channel if and only if they use the same antenna port [13].

In practice, the antenna port can be said to be defined by the transmitted reference signal. The reference signal could have been transmitted from a single physical antenna element or using a beamformer applied on a subarray of elements. For example, even if two signals are transmitted using the same physical antennas they will correspond to different antenna ports if they are beam-formed with different weights, since the corresponding effective channels will be different.[21] The receiver can use a reference signal transmitted on an antenna port to estimate the channel for this antenna port and this channel estimate can subsequently be used for decoding data transmitted on the same antenna port. For example, the DM-RS (see Section 2.5) in LTE and NR can be used for channel estimation to decode data transmitted on the same antenna port.

QCL is defined in a similar manner: "Two antenna ports are said to be quasi-co-located if properties of the channel over which a symbol on one antenna port is conveyed can be inferred from the channel over which a symbol on the other antenna port is conveyed". The main difference between the antenna port and QCL definitions is that the former speaks of *the channel* while the latter refers to the *properties of the channel*. Thus QCL is a less stringent requirement than an antenna port since only the properties of the channel and not the channel itself need to be the same for quasi-co-located antenna ports. If two signals have been transmitted on two closely spaced, but different, antennas, they could experience different channels due to fading but the large-scale properties of the two channels will probably be the same. In such a case, the two antennas would be different antenna ports but they would be quasi-collocated. Large-scale properties include second-order statistics of the channel such as delay/Doppler spread, average channel gain, etc. Such information can, for example, be useful to

[21] If the beam-forming weights are known to the receiver, it could be the same antenna ports, since then the beam-forming weights need not be considered as part of the channel.

the UE for performing channel estimation. An example of antenna ports not being QCL is if they use antennas from different locations, for example in multipoint transmission.

The concept of QCL was introduced in LTE Release 11 to support different types of multipoint transmissions. In NR, QCL has a more central role to play since, e.g., the beam management procedures rely heavily on the QCL concept. In particular, spatial QCL assumptions are used to help the receiver to select an analog Rx beam during beam management; see Section 7.3.4.4 for details.

7.2.6 CSI ACQUISITION

Acquiring CSI is one of the most important aspects of multiantenna techniques since the quality of the CSI is often the limiting factor on the performance of multiantenna techniques. CSI in general can be detailed such as the complex channel matrix for every subcarrier in an OFDM system or coarse, such as the direction to a UE in LoS. Advanced multiantenna techniques such as MU-MIMO and interference suppression exploit detailed channel knowledge and therefore put high demands on the CSI acquisition. The performance gain with such techniques can be very high if perfect CSI is available but in practice it must be estimated based on measurements. If a sufficient CSI quality cannot be obtained, the gain with advanced techniques that rely on detailed CSI may vanish.

In 3GPP, the term 'CSI' has a more specific meaning, namely the particular reports from the UE to the BS indicating the channel quality and other channel properties. Most of the CSI parameters in such reports are actually preferred downlink transmission parameters rather than explicit channel parameters. The UE estimates the channel quality and other properties based on measurements on reference signals transmitted in the downlink. The chief downlink reference signal in LTE and NR for computing CSI is the CSI-RS; see Section 2.5. A CSI report can be periodic, semi-persistent, or aperiodic. Periodic reports occur at regular time instants, while aperiodic reports are triggered on a per need basis. Semi-persistent reports are transmitted periodically until further notice. Aperiodic reports are generally more detailed than periodic reports. In NR, CSI consists of the following components:

- The channel quality indicator (CQI). CQI is an index to the highest modulation and coding scheme (MCS) that would result in a block-error probability of at most 10%, conditioned on a certain transmission hypothesis.
- The rank indicator (RI). RI is a recommendation of which transmission rank to use in codebook-based precoding.
- The precoding matrix indicator (PMI). PMI indicates the preferred precoder to use in codebook-based transmission, conditioned on the indicated transmission rank.
- The CSI-RS resource indicator (CRI). CRI is used for indicating the best beam in beam management or beam-formed CSI-RS transmissions. This is described later in this chapter.
- The layer-1 reference signal received power (L1-RSRP). L1-RSRP is the received power as measured by the UE on a configured reference signal, e.g., a CSI-RS. Layer-1 (physical layer) is in contrast to layer-3 (RRC) RSRP for which additional filtering is applied.
- The strongest layer indicator (SLI). SLI indicates which column in the precoding matrix that corresponds to the layer with the highest SINR. This can be used for transmitting PT-RS on the strongest layer to achieve the most accurate phase tracking at millimeter-wave frequencies.

In this section CSI is used in a broader meaning and not necessarily limited to the parameters listed above.

CSI is needed for several purposes such as scheduling, selection of multiantenna scheme, rank and link adaptation (setting MCS), coherent demodulation, and for determining precoding or combining weights in multiantenna transmission and reception. CSI acquisition is facilitated by transmitting predefined reference signals known to the receiver. The receiver can then estimate the channel by correlating the received signal with the corresponding reference signal. Channel state information at the receiver (CSIR) is thereby relatively straightforward to acquire. Acquiring CSIT is more challenging. The two main alternatives for acquiring CSIT are feedback and reciprocity based. In feedback-based CSI acquisition, CSI is obtained by feedback from a receiving node that has performed channel estimation on the transmitted reference signals. Reciprocity-based CSI acquisition relies on the assumption that CSIT can be obtained from measurements on received reference signals. In this section we discuss CSI acquisition aspects with focus on acquiring CSIT at the BS for downlink transmission.

7.2.6.1 Reciprocity Based

Reciprocity may in broad terms be defined as that knowledge about the Tx channel can be inferred from knowledge about the Rx channel. One can think of different degrees of reciprocity, such as:

- The complex channel matrix is the same for the Tx and Rx channels. This is the strongest form of reciprocity and is in general only possible to achieve in TDD operation.
- The channel second-order statistics, e.g., covariance matrix, is similar for Tx and Rx. Wideband and long-term channel properties are similar for the uplink and downlink also for FDD systems which can be utilized for reciprocity-based operation. For example, precoding based on channel covariance can to some extent be based on reciprocity.
- The angles of departures are the same as angle of arrivals. Fast fading is caused by complex superposition of multiple channel rays, and they will sum up differently for the DL and UL carriers in an FDD system. The directions of the rays are, however, typically not dependent on the carrier frequency, so they can be reciprocal in FDD. For example, beam-forming in a LoS channel can be based on reciprocity.

The rest of this subsection is concerned with the strongest form of reciprocity described in the first bullet above. Reciprocity-based CSIT acquisition for downlink transmission can be performed by configuring the UE to transmit SRSs in the uplink. The BS can perform channel estimation on the received SRSs to obtain an estimate of the uplink channel matrix. If reciprocity holds, the uplink channel estimate can then be used as an estimate of the downlink channel. A clear advantage with reciprocity-based CSI acquisition compared to feedback based acquisition is a reduced overhead, since no feedback signaling is needed to obtain CSIT. Also the reference signaling overhead is reduced if the number of BS antenna elements is larger than the number of UE antenna elements, which typically is the case. With feedback-based CSI acquisition, one reference signal per BS Tx antenna element needs to be transmitted, while one reference signal per UE Tx antenna element needs to be transmitted with reciprocity-based CSI acquisition. Another advantage with reciprocity-based CSI acquisition is that it is less reliant on specification support than feedback-based methods in the sense that it can be used for any number of BS antenna elements. Feedback-based methods typically use some kind of quantized channel feedback such as codebooks, and these need to be specified for each supported number of BS antenna ports and need to be known to both the BS and the UE.

A disadvantage with reciprocity-based methods is that they can only be used in TDD systems if the strongest form of reciprocity should be utilized. However, even if the system operates in TDD, sometimes only partial reciprocity can be guaranteed. Although the propagation channel is reciprocal in a TDD system,[22] the BS transceivers may not be. To utilize reciprocity, also the BS transceivers need to be reciprocal, i.e., the Tx and Rx branches need to have the same characteristics. This can be achieved to a certain level of accuracy by reciprocity calibration of the transceivers [50]. Passive components like the antenna radiating elements are typically reciprocal as long as the same elements are used for transmission and reception, which, however, may not always be the case.

Another type of partial reciprocity is when the UE has fewer Tx than Rx branches, which is quite common in hand-held devices in order to save battery time. In this case the UE cannot sound all antennas simultaneously, since there are not enough Tx branches. This implies that not the full channel matrix can be estimated from uplink sounding, only the channels to the antennas which have a connected Tx branch. A possible solution to this problem is to switch the Tx branches between the antennas sequentially in time until all antennas have been sounded. Partial reciprocity may also be due to the number of carriers being different in uplink and downlink, which could be the case in carrier aggregation. Reciprocity can then only be utilized for the carriers that are common for uplink and downlink.

The discussion of reciprocity has so far concerned the channel between a BS and the served UE, i.e., the desired link. Interfering links are also important to take into account in multiantenna transmission. Interference can have a significant impact on link adaptation, scheduling, and precoder design. Even if the propagation is reciprocal, the interference is typically different in uplink and downlink. For example, the downlink interference received by a UE from another, non-serving, BS cannot be measured in uplink by the serving BS. Therefore, even if reciprocity holds for the desired link in a TDD system, it needs to be complemented by feedback containing information regarding the interference. The interference feedback does not have to be the interference in itself; it could be the impact of the interference, e.g., on the quality of the signal after the receiver. Since the BS may not know which interference suppression capability the UE has in its receiver, it can be better to feed back an SINR related quantity instead of the actual interference level. One such quantity that is used in LTE and NR is CQI.

A difference between the downlink and uplink, regardless of TDD or FDD, is that the UE usually has a much lower Tx power than the BS. This may cause problems with reciprocity-based CSI acquisition in some scenarios if the UE Tx power is not high enough to yield sufficient SNR in the channel estimation performed by the BS. Poor channel estimation quality will lead to poor performance of downlink multiantenna transmission schemes that rely on detailed channel knowledge, e.g., MU-MIMO. It may then be better to use feedback-based CSI acquisition, since the UE can perform channel estimation at a higher SNR than the BS due to the higher Tx power in the BS.

7.2.6.2 Feedback Based

In downlink feedback-based CSIT acquisition the BS transmits reference signals that are known to the UE. The UE estimates the downlink channel and feeds back a quantized channel estimate by uplink

[22] Propagation is reciprocal if the medium is linear. Furthermore, the uplink channel estimation and downlink transmission must be performed within the channel coherence time so that the channel does not change due to fading.

signaling. The feedback can be explicit or implicit. With explicit feedback, a quantized and possibly compressed representation of the channel could be reported to the BS, e.g., a quantization of the channel matrix itself or of the principal eigenvectors of the channel covariance matrix. The channel feedback can also be complemented by feedback of the interference experienced by the UE. With implicit feedback, preferred transmission parameters are fed back. One example of implicit feedback is codebook-based CSI acquisition where a number of candidate precoding matrices are collected in a codebook. The UE then evaluates, based on its channel estimates, which precoding matrix in the codebook would give the highest performance if used by the BS [29]. The UE then feeds back an index to the precoder matrix in the codebook, in LTE and NR called a PMI.

Note that the use of codebooks for CSI acquisition does not mean that the BS has to use a precoder from the codebook in the data transmission. This depends on whether the reference signals used for coherent demodulation are precoded in the same way as data or not. If the demodulation reference signals are precoded with the same precoder as the data, the UE does not need to know which precoder the BS has used in order to demodulate the data. The codebook is in this case only used to feed back CSI to the BS. The BS can then use this information to design an arbitrary precoder without informing the UE. An example of a transmission scheme where the demodulation reference signals are precoded in the same way as data is transmission mode (TM) 9 in LTE. If instead the reference signals for coherent demodulation are transmitted per antenna element without any precoding, the UE needs to know which precoder the BS has applied to the data transmission so that the UE can apply this precoder to the estimated channel before demodulating the data. An example of such a transmission scheme is TM 4 in LTE, where non-precoded reference signals, so-called cell-specific reference signals (CRSs) are used for demodulation.

The main advantage of feedback-based CSI acquisition is that it does not rely on reciprocity and can thus be used for both FDD and TDD. A disadvantage is the signaling overhead required to feed back the CSI, especially if high-resolution CSI is needed. This can be prohibitive for multiantenna transmission schemes that need CSI with high resolution, e.g., MU-MIMO, if the number of antenna ports is large. Feedback-based CSI acquisition is also closely tied to the standard, since it must be specified how the channel should be represented. For example, with codebook-based precoding a codebook for each supported number of antenna ports needs to be specified.

A way to reduce the feedback overhead and CSI computation complexity with codebook-based CSI is to design the codebook based on prior knowledge about a certain channel structure. For example, if the channel is expected to be correlated between antenna elements a smaller codebook can be used than if the channel is uncorrelated, since then the channel coefficient for one antenna element is related to the others. A high correlation between antenna elements is a reasonable assumption for closely spaced, co-polarized antenna elements in a channel with limited angular spread, e.g., when using a tower-mounted macro BS antenna. Several codebooks in the LTE and NR specifications have been designed based on this assumption.

To appreciate the construction of such a codebook, consider the array response vector $\mathbf{a}(\theta)$ for a ULA with N elements when a single plane wave from direction θ relative to the array boresight is impinging upon the array. This vector can be written as

$$\mathbf{a}(\theta) = \begin{bmatrix} 1 & e^{-j2\pi \frac{d}{\lambda} \sin\theta} & \dots & e^{-j2\pi (N-1) \frac{d}{\lambda} \sin\theta} \end{bmatrix}^T \tag{7.36}$$

where λ is the carrier wavelength and d is the distance between two adjacent elements. Making the substitution $\psi = d\sin(\theta)/\lambda$ we obtain

$$\mathbf{a}(\psi) = \begin{bmatrix} 1 & e^{-j2\pi\psi} & \dots & e^{-j2\pi(N-1)\psi} \end{bmatrix}^T. \tag{7.37}$$

Hence, $\mathbf{a}(\psi)$ has the same structure as the vectors in a discrete Fourier transform (DFT) matrix. Beam-forming in a direction θ can be achieved by applying a weight vector $\mathbf{w} = \mathbf{a}^*$, where * denotes complex conjugate, to the array. A codebook designed for a ULA can therefore be constructed by DFT vectors corresponding to beam-forming in a number of hypothesized directions, resulting in a so-called DFT codebook. A set of beams corresponding to beam-formers in different directions is often called a grid-of-beams (GoB).[23]

DFT codebooks are used extensively in LTE and NR for codebook-based CSI acquisition. For a ULA with N elements, the kth precoding vector in such a codebook is given by

$$\mathbf{w}_{\text{ULA}}(k) = \frac{1}{\sqrt{N}}\begin{bmatrix} 1 & e^{j2\pi\frac{k}{QN}} & e^{j2\pi(N-1)\frac{k}{QN}} \end{bmatrix}^T,\ k = 0, 1, \dots, QN-1. \tag{7.38}$$

Here, Q is a so-called oversampling factor which is used to obtain a finer granularity in the angular domain than what a codebook with N orthogonal DFT vectors would provide. For a UPA, a corresponding precoding vector is obtained by the Kronecker product of the 1-D precoding vectors in each dimension, i.e., $\mathbf{w}_{\text{UPA}}(k,l) = \mathbf{w}_{\text{ULA}}(k) \otimes \mathbf{w}_{\text{ULA}}(l)$, where $\otimes$ denotes the Kronecker product. A common BS antenna array architecture is the UPA with dual-polarized elements. Since orthogonal polarizations typically fade independently, a reasonable approach is to have DFT vectors per polarization and a phase difference between the DFT vectors for different polarizations. Such codebooks have been constructed for both LTE and NR and contain vectors of the form

$$\tilde{\mathbf{w}}_{\text{UPA}}(k,l,\phi) = \begin{bmatrix} \mathbf{w}_{\text{UPA}}^T(k,l) & e^{j\phi}\mathbf{w}_{\text{UPA}}^T(k,l) \end{bmatrix}^T. \tag{7.39}$$

DFT-based codebooks are also used when transmitting multiple precoded layers in spatial multiplexing. The codebooks therefore contain different combinations of DFT vectors with one vector for each layer. An entry in the codebook that corresponds to multilayer transmission is therefore a matrix. When the BS has received a PMI from a UE it can use the corresponding precoding vector for subsequent data transmissions. Alternatively, it can interpret the PMI as a quantized representation of the channel and design another precoder based on this and possibly also other information.

The feedback overhead associated with codebook-based CSI acquisition can be reduced further by exploiting that some channel properties are frequency selective, while others may be the same over the system bandwidth. For example, directional properties are typically the same over the bandwidth, while polarization properties are frequency selective. By separating the codebook into a wideband and a frequency-selective part, the feedback overhead can be reduced since the two parts can be reported with different frequency granularity. An example of this is the so-called dual-stage codebook introduced in LTE Release 10, in which a precoding matrix in the codebook is factorized as $\mathbf{W} = \mathbf{W}_1\mathbf{W}_2$, where

[23] The beam-forming vectors in a GoB do not necessarily have to be DFT vectors.

$\mathbf{W}_1$ captures the long-term/wideband properties and $\mathbf{W}_2$ the short-term/frequency-selective part. Such a codebook is useful for dual-polarized arrays with closely spaced antenna elements. In this case, $\mathbf{W}_1$ can perform beam-forming over co-polarized antenna elements while $\mathbf{W}_2$ performs co-phasing of polarizations.

Advances in active array antenna technology leads to the circumstance that the number of digital elements in a BS array antenna is continuously increasing. Obtaining CSI for a large number of antenna ports based on codebooks can be prohibitive if this number becomes too large, since it leads to a large overhead and CSI computational complexity. It will also lead to a lower antenna gain for each antenna port if the port now becomes connected to a single antenna element instead of a subarray of elements. This in turn implies a lower SNR in the channel estimation and thereby a lower CSI quality.

A remedy to these problems is to dynamically beam-form the reference signals in a UE-specific manner. In this way the channel is estimated in beam space instead of element space. This leads to a lower required number of reference signals and also a higher antenna gain for the transmitted reference signals. Reference signals can then be transmitted in a few beams instead of on many antenna elements. The UE then measures the channel quality on each beam-formed reference signal and reports the preferred beam and CSI for the selected beam. However, this approach requires some prior knowledge about the direction to a UE. This can be obtained by, e.g., uplink measurements or from previous downlink reference signal transmissions. Since only directions are of interest, full reciprocity is not necessary, so uplink measurements can be used also for FDD. Previous reference signal transmissions could be transmitted with a low density in time and/or frequency to give coarse estimates about directions without incurring too much overhead. More dense beam-formed reference signals can then be transmitted in these directions to acquire detailed channel knowledge in this angular region. A beam tracking procedure for moving UEs can also be performed where a few candidate beams centered around the active beam are monitored by the UE.

Beam-forming of reference signals can be beneficial when the BS has many antenna elements and when the number of UEs in the cell is low. An advantage for the UE is reduced complexity in the CSI calculations since only a few beams have to be evaluated instead of many precoder candidates. An example of CSI acquisition using beam-formed reference signals is beam-formed CSI-RS introduced in LTE Release 13, also called CSI feedback class B [13].

7.2.7 MASSIVE MIMO

Massive MIMO has received considerable attention in both the academic research and the wireless industry in recent years and is a key technology component for 5G, both for NR and the evolution of LTE. The term has a relatively well-defined meaning in academic circles, while the wireless industry often uses it in broader terms. In academic parlance, massive MIMO is usually associated with a digital array having a large number (hundreds) of antenna elements serving much fewer (tens) UEs using reciprocity-based MU-MIMO. It is often considered to be restricted to TDD since obtaining CSIT in an FDD system is difficult if the number of antenna elements is large. In the wireless industry, the term ‘massive MIMO’ is often used for large antenna arrays with somewhat less number of elements which are not necessarily used for MU-MIMO. The industry sometimes also uses the term ‘massive MIMO’ for analog beam-forming with many antenna elements at millimeter-wave frequencies.

Massive MIMO was initially introduced as an asymptotic notion of letting the number of base station antennas grow to infinity [31]. This leads to several interesting theoretical results, such as that

simple linear precoding with MRT is optimal and that the impact of fast fading, noise, and some types of interference and hardware impairments vanish thanks to averaging over infinitely many antennas. In order not to clutter the entire radio resource grid with pilots, channel reciprocity is usually assumed, so that the downlink channel state can be inferred from uplink measurements. In theory, this could enable a very high capacity with simple transceivers and scheduling strategies. Of course, in any real-world implementation the number of antenna elements has to be finite. However, massive MIMO does represent a paradigm shift from using a few high-end radio transceivers in traditional base stations to a large number of transceivers with relaxed quality requirements per transceiver. Indeed, several field trials from both academia and the industry have shown that impressive spectral efficiencies can be achieved with massive MIMO; see, e.g., [17,36] and Section 1.3.4.

Regardless of how massive MIMO is defined, antenna arrays with a large number of elements are instrumental to fulfilling the performance requirements for 5G, whether they are used for reciprocity-based MU-MIMO to increase capacity at lower frequencies, or for analog beam-forming providing coverage at millimeter-wave frequencies. We do not pursue the massive MIMO-specific discussion any further here, since many of the issues associated with massive MIMO, such as CSI acquisition and precoder design, are treated in other sections of this chapter.

7.3 MULTIANTENNA TECHNIQUES IN NR

For low frequencies, multiantenna techniques in NR build to a large extent upon those in later releases of LTE. However, NR does contain a number of improvements that can increase the capacity and make it easier to adapt to diverse use cases. A major enhancement is a new flexible, modular, and scalable CSI framework. This framework also includes a high-resolution CSI reporting mode targeting improved MU-MIMO operation. For high frequencies, there are more fundamental changes, since LTE was not designed for millimeter-wave spectrum. The main new feature is the support for analog beam-forming in both the BS and the UE. This is called beam management and is described in Section 7.3.4.

The first part of this section discusses MIMO transmission over digital antenna ports and how to acquire CSI to enable such a transmission. The second part is about establishing and maintaining beam pair links for analog beam-forming, i.e., beam management. While the first part is focused on low frequencies and the second part on high frequencies, the techniques can be combined using, e.g., hybrid beam-forming with multipanel arrays. Beam management can then be used to determine the analog beam-formers in the panels and the CSI acquisition and MIMO techniques to determine the digital precoding across panels.

In LTE, the different downlink multiantenna transmission techniques are specified in ten different transmission modes. There are transmission modes for supporting diversity, beam-forming, spatial multiplexing, and CoMP transmission. Beam-forming and spatial multiplexing can be combined using precoded spatial multiplexing transmission modes. LTE supports open- and closed-loop codebook-based precoding as well as non-codebook-based precoding. There is also a transmission mode for MU-MIMO transmission.

The transmission modes also differ in which reference signals are used for demodulation and how CSI is acquired by the UE and fed back to the network. In the early LTE releases, demodulation and CSI acquisition was based mainly on CRS transmissions. These are transmitted on every antenna, in every subframe and resource block and are common for all UEs in a cell. Initially, LTE was designed

for a relatively few number of antennas. As the number of supported antennas was increased in later releases, new reference signals were introduced in order to reduce the large overhead a CRS-based transmission would incur. Separate reference signals for demodulation and CSI acquisition were introduced in Release 10 with DM-RS[24] and CSI-RS.

Reference signals for coherent demodulation typically require higher time-frequency density than what is needed for CSI used for the selection of transmission parameters. By having separate reference signals for demodulation and CSI acquisition the density can be optimized for their respective purpose. Furthermore, reference signals for demodulation need only to be transmitted when there is data to transmit and there is need to occupy only the bandwidth scheduled for the data transmission. Moreover, and perhaps most importantly, the number of DM-RS ports scales with the number of layers rather than the number of antenna elements. By letting DM-RS be a UE-specific signal that is transmitted only when there is data to transmit to the UE, the overhead and intercell interference induced by the reference signals can be reduced compared to the always-on CRS. This is particularly important at low traffic load where CRS interference could limit the data transmission performance. DM-RS can also be precoded with an arbitrary precoder, e.g., with one of the interference suppression MU-MIMO precoders described in Section 7.2.2.2. This is transparent to the UE, so it does not need to know which precoder the BS has applied. Furthermore, CRS is transmitted over the whole bandwidth even if there is no data traffic in the cell, leading to unnecessarily high energy consumption.

For these reasons, CRS has been removed in NR. This provides a more lean, energy-efficient, flexible and scalable design. Furthermore, the different transmission modes have been removed in NR. There is only one downlink data transmission scheme, similar to TM 10 in LTE, in which channel estimation for demodulation and CSI acquisition is based on DM-RS and CSI-RS, respectively. In LTE the different functions of CSI configuration, measurement, reporting, and multiantenna transmission were tightly coupled. This has lead to multiple transmission modes, CSI classes, and configurations and many different options to choose from. In NR, these functions have been decoupled to allow for a more flexible configuration, making it easier to optimize for diverse use cases. This also makes it more scalable and renders the introduction of future enhancements easier. For example, NR provides more flexibility in the reference signal placement and density in the time-frequency grid, dynamic triggering of different types of reports, and fast feedback.

Besides a more flexible CSI acquisition framework, NR also contains enhancements targeting improved MU-MIMO transmission such as increased number of MU-MIMO layers, high-resolution CSI feedback and improved interference measurements. Since TDD is expected to be more common in NR deployments, it also has improved support for reciprocity. This includes increased SRS capacity and coverage by allowing multiple SRS symbols in one slot, SRS switching when a UE has fewer Tx than Rx branches, and so-called non-PMI feedback for improved link adaptation when the BS lacks information regarding the downlink interference (see Section 7.3.1.1).

7.3.1 CSI ACQUISITION

As alluded to previously, the purpose of the new CSI acquisition framework in NR is to decouple different CSI functionalities so that they can be configured independently leading to a modular, flexible,

[24] DM-RS was supported already in the first LTE release but was then only used for single-layer data transmission in TM 7.

and scalable design where it becomes easier to introduce new features and adapt to different use cases. The framework is common for MIMO transmission and beam management.

NR defines one or more report settings and one or more resource settings. A report setting is then linked to one resource setting for channel measurement and one resource setting for interference measurement. A resource setting defines which reference signal resources should be used for measurements, what type of reference signal, and the time-domain behavior of the resource configuration, i.e., if it is periodic, semi-persistent, or aperiodic. Two different types of reference signals are used for CSI acquisition in NR, CSI-RS and the synchronization signal block (SSB). The SSB consists of the PSS, SSS, and physical broadcast channel (PBCH). CSI acquisition using SSB is only used for beam management; see Section 7.3.4. A report setting tells the UE how the reporting shall be performed and what a CSI report shall contain. For example, it can include which CSI parameters to report, how the reporting shall be made in the time and frequency domain, e.g., periodic/semi-persistent/aperiodic and wideband/subband, measurement restrictions, codebook configuration, etc. The network can dynamically select one or multiple report settings to trigger the desired CSI reports.

Different transmission schemes and use cases can have different CSI requirements. SU-MIMO transmission typically has lower requirements on the CSI resolution than MU-MIMO, since there are better abilities to suppress SU-MIMO interlayer interference in the UE receiver. This is due to the fact that the number of transmitted SU-MIMO layers cannot be larger than the number of Rx antennas in the UE. In MU-MIMO transmission the total number of layers can be larger than the number of Rx antennas in the UE, which makes it more difficult to suppress interference in the receiver. Therefore, MU-MIMO transmission needs to some extent to rely on transmitter interference suppression, e.g., ZF or MMSE precoding, and this requires high-resolution CSI in order to be effective.

For downlink MIMO transmission, NR has two different CSI reporting types in order to support different CSI requirements. These are called Type I and II, respectively. Type I is a moderate resolution reporting mode targeting SU-MIMO operation while Type II has higher CSI resolution aimed at supporting MU-MIMO transmission, at the expense of a higher feedback overhead. Type I CSI gives information only about the strongest channel direction, while Type II captures channel multipath by representing the channel as a linear combination of multiple orthogonal DFT vectors. Type II is suitable for FDD since detailed CSI cannot be obtained by reciprocity in this case. It can also be useful for TDD if full reciprocity cannot be achieved, e.g., due to uncalibrated transceivers.

Feedback-based CSI in NR is based on codebooks. A number of codebooks have been defined, supporting up to 32 antenna ports. They all have a similar structure and are based on a dual-stage codebook where a precoder matrix is factorized into two components, $\mathbf{W} = \mathbf{W}_1\mathbf{W}_2$, where $\mathbf{W}_1$ captures the wideband channel properties and $\mathbf{W}_2$ the frequency-selective part. An example of a wideband channel property is the dominant propagation directions in the channel, while frequency-selective fading obviously is a frequency-selective property. The codebooks have been designed for dual-polarized UPAs for which the dual-stage codebook is a suitable design. For such arrays, dominant channel directions are conveniently modeled by 2-D DFT vectors for each polarization. The wideband matrix $\mathbf{W}_1$ is therefore composed of one or multiple such DFT vectors. Applying $\mathbf{W}_1$ as a precoder matrix can then be interpreted as transmitting with beams pointing in the dominant channel directions by using the corresponding DFT vectors as beam-forming weight vectors. The frequency-selective part of the dual-stage codebook, $\mathbf{W}_2$, accounts for co-phasing of polarizations, beams, and/or panels.

Type I CSI reporting consists of a single-panel and a multi-panel codebook. The single-panel codebook is similar to the full-dimension MIMO (FD-MIMO) codebooks in LTE Release 13 [13] and

supports transmission ranks up to eight. For rank one, it amounts to selecting a single beam from an oversampled grid of DFT beams and co-phasing of the polarizations. The DFT vector corresponding to the selected beam is then contained in $\mathbf{W}_1$, while the co-phasing factors are in $\mathbf{W}_2$. The multi-panel codebook is an extension of the single-panel codebook by including co-phasing of panels in a multi-panel array. Up to four panels are supported. The co-phasing can be either wideband or frequency selective.

In Type II CSI reporting, $\mathbf{W}_1$ contains DFT beam-forming vectors for multiple selected beams and $\mathbf{W}_2$ is a beam combining matrix. Up to four beam-forming vectors can be selected for $\mathbf{W}_1$ from a set of orthogonal DFT vectors. The set of orthogonal DFT vectors can be rotated so that the corresponding beams are better aligned with the dominant directions in the channel. The $\mathbf{W}_2$ matrix contains frequency-selective co-phasing factors for combining the beams selected in $\mathbf{W}_1$. There is also an amplitude scaling factor for each beam which can be wideband or a combination of wideband and frequency selective. The beam selection for $\mathbf{W}_1$ is common for both polarizations and, in the case of rank two, both layers while the co-phasing and amplitude scaling is selected independently per polarization and layer. The precoding vector for polarization r, layer l, and a particular frequency subband is thus represented as

$$\mathbf{w}_{r,l} = \sum_{k=1}^{K} \mathbf{a}_k b_{r,l,k} e^{j\phi_{r,l,k}} \tag{7.40}$$

where the $\mathbf{a}_k$ are the selected beam-forming vectors, $b_{r,l,k}$ is the amplitude scaling factor, and $\phi_{r,l,k}$ the phase. The $\mathbf{a}_k$ are the same for all subbands, $\phi_{r,l,k}$ is selected for each subband, and the selection of $b_{r,l,k}$ is either wideband or a combination of per subband and wideband. In this way, channel multipath can be conveyed by the CSI report, which makes MU-MIMO precoding with interference suppression more effective. Construction of a precoding beam using Type II CSI reporting is illustrated in Fig. 7.10.

The Type II codebook supports only transmission ranks one and two in order to keep the feedback overhead at a reasonable level. Due to the relatively high feedback overhead of Type II reporting it is suitable for low mobility UEs. Tracking a high-mobility UE with high spatial resolution would also require a high CSI-RS density. Type I reporting may therefore be a better choice for high-mobility UEs. The Type II beam combining codebook is intended to be used with non-beam-formed CSI-RS. Type II also has a port selection codebook for beam-formed CSI-RS. The $\mathbf{W}_1$ matrix is then a port selection matrix and amplitude scaling and co-phasing is performed in the same way as for the beam combining codebook.

Although the CSI report contains indices to parameterized precoding vectors, the BS can use any precoder in the data transmission since the DM-RS is precoded in the same way as data. The CSI report is only a recommendation of transmission parameters which the BS can choose to follow or not. The CSI report can be used by the BS to reconstruct the vector in (7.40), which can be interpreted as an approximation of the channel as estimated by the UE. Based on this, the BS can for example apply one of the interference suppression MU-MIMO precoders described in Section 7.2.2.2.

The Type II codebook is similar to the LTE Release 14 advanced CSI codebook but allows for selection of more beams and has finer granularity in the amplitude and phase quantization. Furthermore, for arrays with more than 16 antenna ports there is a restriction in LTE that beam selection can be made only from a subset of all possible beams, a restriction that has been removed in NR. These enhancements can give a substantial performance gain of the NR codebook compared to the LTE codebook. Simulation results in [38] for MU-MIMO transmission in the 3GPP UMi scenario showed 24%

and 56% increased average and cell edge user throughput, respectively, of the NR Type II codebook over the LTE Release 14 advanced CSI codebook. A Type II codebook for 32 BS Tx antennas using four beams and combined wideband and subband feedback of the amplitude scaling was used in the simulations.

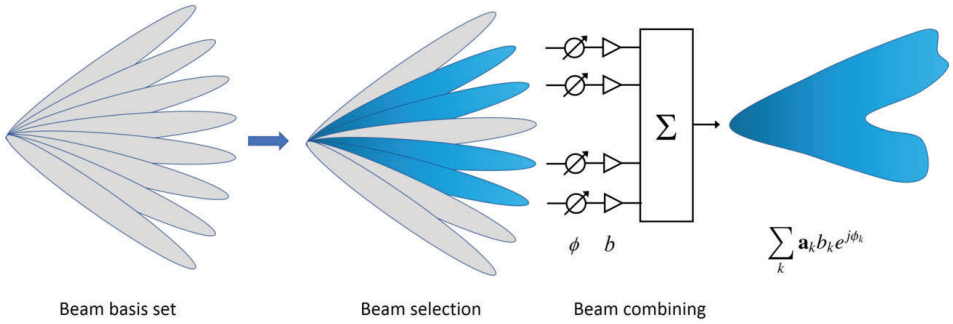


FIGURE 7.10

Type II CSI. Illustration of CSI Type II codebook.

7.3.1.1 Interference Measurements

The discussion of CSI has so far concerned the channel to the served UEs. Another important CSI component is the interference experienced by the UE. In the early LTE releases it was not specified how the UE should estimate the interference; it was up to UE implementation [13]. A typical approach was to estimate the intercell interference based on CRSs. A problem with this approach is that CRS is always transmitted regardless of the traffic load. At low load the intercell interference will therefore be dominated by CRS transmissions from neighboring cells. This means that CRS-based interference estimation will overestimate the interference on the data channel at low traffic load. Another problem is that the network has no control over how the UE averages its interference estimates.

To improve the interference estimation, a so-called CSI-IM configuration was introduced in LTE Release 11 [13]. This allows the network to control in which resource elements the UE should measure interference. A CSI-IM resource is just a CSI-RS resource in which nothing is transmitted from the BS to the UE that has been configured with the resource, a so-called ZP CSI-RS. Since the CSI-IM collides with data transmissions from other cells instead of CRSs, interference estimation at low traffic load can be improved.

In NR, the bandwidth of CSI-IM is configurable so that the UE can be configured to measure interference on only a part of the frequency band. This can be useful when mixing different services in different parts of the frequency band, so that interference is measured only in the part that is occupied by the service that the UE uses.

In LTE, CSI-IM is used for intercell interference estimation in SU-MIMO operation. For MU-MIMO it is important to also consider multiuser interference between the co-scheduled UEs within the cell. NR has therefore introduced support for multiuser interference measurements based on non-zero-power CSI-RS (NZP CSI-RS). Each UE in a multiuser group is then configured with one NZP CSI-RS resource for channel measurement and one NZP CSI-RS resource per co-scheduled UE for multiuser

interference measurements. CSI-IM with ZP CSI-RS can still be used for intercell interference estimation in NR.

With reciprocity-based CSI acquisition, the downlink channel can be estimated from uplink sounding. However, in order to choose proper transmission rank and link adaptation parameters, the BS needs to know the SINR in the UE, which also depends on the interference. To convey this information to the BS for reciprocity operation, NR supports so-called non-PMI feedback where the UE reports RI and CQI, but not PMI. The BS can estimate the downlink channel based on SRS transmission from the UE and design a precoding matrix based on the channel estimate. The BS then transmits precoded CSI-RS with the determined precoding matrix and the UE can calculate RI and CQI based on the precoded CSI-RS taking its particular receiver algorithm into account. If the BS performs subsequent data transmission with the same precoding matrix, there is no mismatch between the data precoding and the precoding assumed in the CSI calculation.

7.3.2 DOWNLINK MIMO TRANSMISSION

The downlink MIMO transmission in NR is mainly based on non-codebook-based precoding using CSI-RS for CSI acquisition and DM-RS for coherent demodulation. For reciprocity-based operation, CSI can be obtained by uplink sounding complemented by RI and CQI feedback. For feedback-based CSI, the codebooks described in the previous section can be used. In this case, a typical downlink MIMO transmission consists of the BS first transmitting CSI-RS for CSI acquisition. The CSI-RS could be beam-formed or transmitted per antenna element. The UE estimates the channel and calculates CSI based on the channel estimate. The CSI can consist of RI, PMI, and CQI. For beam-formed CSI-RS, it can also include CRI to indicate the best beam. The UE then sends a CSI report to the BS which uses the report to determine MIMO transmission parameters and performs the data transmission. NR supports MU-MIMO transmission with up to 12 layers with orthogonal DM-RS ports. Up to eight UEs can be spatially multiplexed. In SU-MIMO transmission, a UE can receive up to eight layers. Hybrid beam-forming with multipanel arrays can be performed by using beam management for the analog beam-forming and the multipanel codebook for the digital precoding.

While LTE supports various Tx diversity schemes such as SFBC, FSTD, and large-delay CDD, Tx diversity is currently not explicitly supported in NR and has to be performed in a specification-transparent manner, e.g., using precoder cycling in frequency.

7.3.3 UPLINK MIMO TRANSMISSION

NR Release 15 supports two transmission schemes for the uplink data channel, i.e., physical uplink shared channel (PUSCH): codebook-based transmission and non-codebook-based transmission. Tx diversity for PUSCH is not specified and thus needs to rely on specification-transparent methods. For DFTS-OFDM codebook-based uplink transmission is limited to rank one, since the use case for DFTS-OFDM is coverage limited UEs. For CP-OFDM, codebook-based uplink transmission supports up to rank four.

Codebook-based transmission can be used for both FDD and TDD. Since it is based on CSI feedback from the BS, it can be used when reciprocity does not hold, e.g., for FDD or for TDD when a UE is not reciprocity-calibrated. In codebook-based uplink transmission for NR, the UE transmits non-precoded SRSs. A UE can transmit one or two SRS resources and an SRS resource can have up to

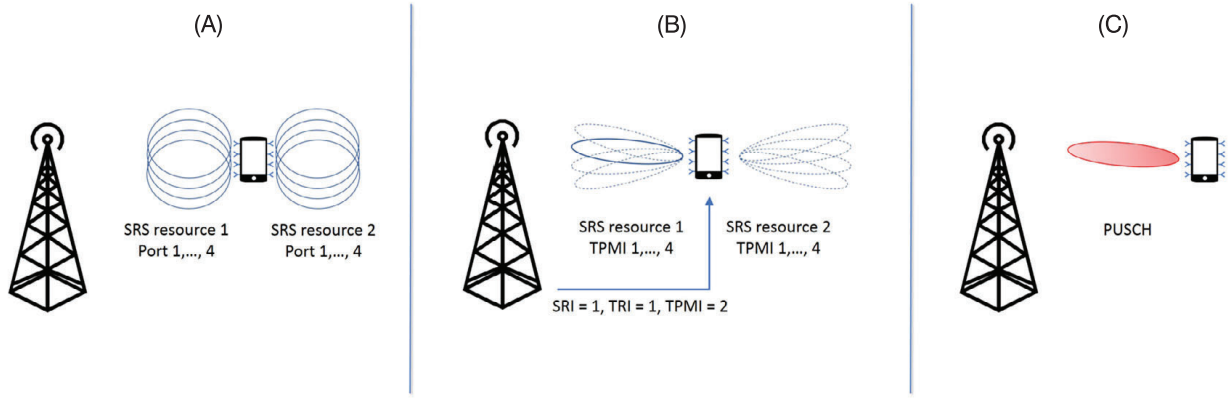


FIGURE 7.11

Codebook-based uplink transmission. An example of codebook-based uplink transmission.

four ports. The BS determines CSI based on the received SRSs and instructs the UE to use the determined CSI parameters. In this case CSI consists of SRS resource indicator (SRI),[25] transmit precoder matrix indicator (TPMI), and transmit rank indicator (TRI). SRI indicates the selected SRS resource, TRI the preferred transmission rank, and TPMI the preferred precoder over the ports in the selected resource, where the precoder is selected from the uplink codebook. The UE then performs the uplink transmission based on the CSI report from the BS.

An example illustrating codebook-based precoding is shown in Fig. 7.11. In this example, the UE has two linear arrays with four antennas each on opposite sides of the device. An SRS resource can in this example correspond to one array and the ports within an SRS resource to the antenna elements within an array. In step (A) in Fig. 7.11, the UE transmits SRS resource 1 on the left array and SRS resource 2 on the right array. The ports within a resource are mapped to the antenna elements in each array. In step (B), the BS evaluates the different precoding matrices in the codebook for the two SRS resources and determines the best SRS resource, transmission rank, and precoding matrix and signals SRI, TRI, and TPMI to the UE. In this example, SRI $=$ 1, TRI $=$ 1, and TPMI $=$ 2. The dashed beams in (B) illustrate all beams in the codebook and the solid beam illustrates the selected beam corresponding to SRI $=$ 1 and TPMI $=$ 2. Finally, in step (C), the UE uses the selected precoding matrix for the PUSCH transmission.

For codebook-based transmission, NR supports three levels of coherence capabilities of a UE: full, partial, and non-coherent. A UE with full coherence can transmit coherently over all antenna ports, i.e., it can control the relative phase between all Tx chains. A UE with partial coherence is able to transmit coherently over pairs of antenna ports, and a non-coherent UE cannot transmit coherently over any antenna ports. The uplink codebook consists of different parts adapted to the different UE coherence capabilities. The part for full coherence is similar to the NR downlink codebook and the part for partial coherence consists of precoders that combine ports within coherent pairs. For non-coherent UEs, a part with one layer per antenna port is used. For example, rank one transmission in this case uses a port

[25] SRI is not reported if only a single SRS resource has been configured.

selection codebook. The BS can configure the UE to use a subset of the entire codebook depending on its coherence capability. The codebook subset can be "full + partial + non-coherent", "partial + non-coherent", or "non-coherent". For example, a fully coherent UE is allowed to use the entire uplink codebook, while a non-coherent UE can only use the non-coherent part.

When reciprocity holds, non-codebook-based transmission is an option. The SRS transmission can then be precoded. The BS can transmit CSI-RS to help the UE to design suitable precoders for the SRS transmission. In this case the UE can transmit up to four SRS resources where each resource has one port. The BS then determines one or multiple SRIs based on the received SRSs and the UE transmits one layer per SRI after it has received the report from the BS. Hence, in this case the transmission rank is equal to the number of SRIs. One way to use non-codebook-based transmission is that the UE designs precoders using channel estimates based on received CSI-RSs and transmits precoded SRS using these precoders; one SRS resource per precoder. The BS then selects one or multiple precoders and indicates these to the UE by reporting the corresponding SRIs. The UE then transmits one layer per indicated SRI using the corresponding precoder. An example that illustrates this is shown in Fig. 7.12. In step (A), the BS transmits CSI-RS from which the UE can estimate the channel. Based on this estimate, the UE designs precoders for the SRS transmission. In step (B), the UE transmits two precoded SRS resources, each with a single port. In step (C), the BS determines the best SRS resource and signals this to the UE with an SRI, in this case SRI = 1. In step (D), the UE performs data transmission using the same precoder as was used for the SRS resource indicated by the BS.

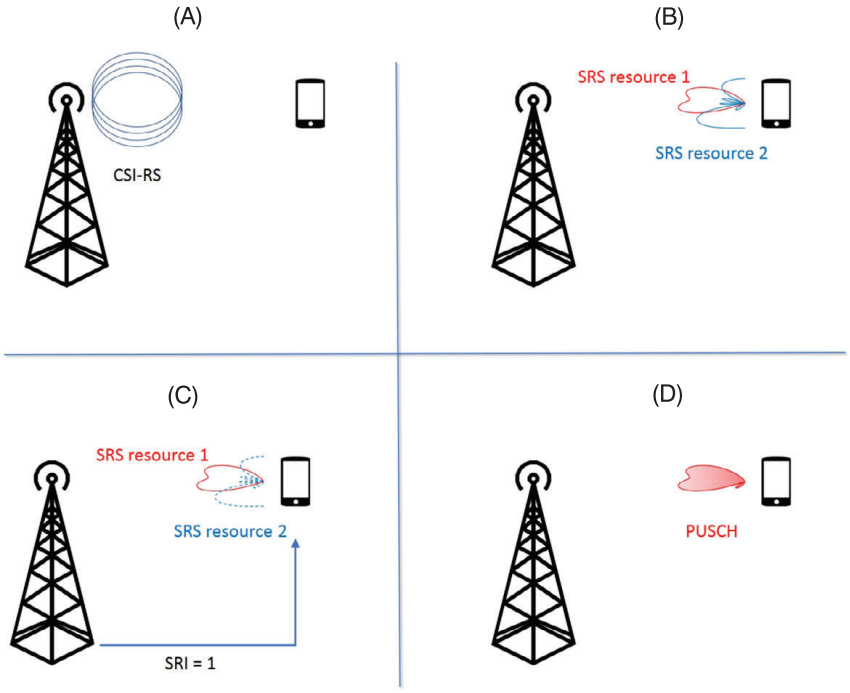


FIGURE 7.12

Non-codebook-based uplink transmission. An example of non-codebook-based uplink transmission.