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

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 ≤ 6 GHz and high frequencies will mean carrier frequencies ≥ 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].

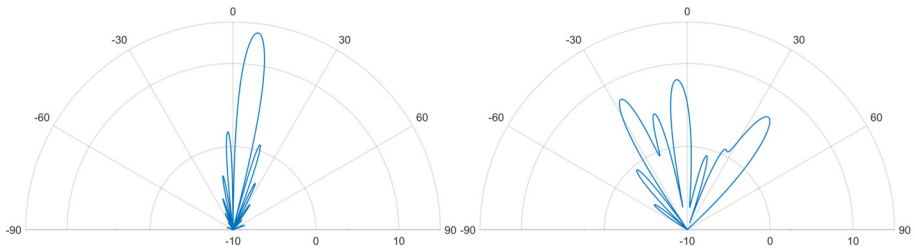

FIGURE 7.1

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.

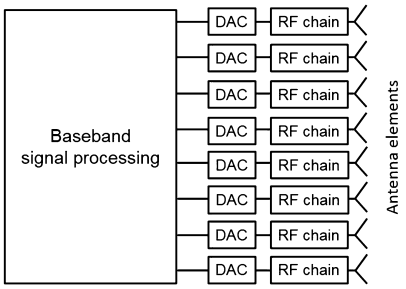

FIGURE 7.2

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.

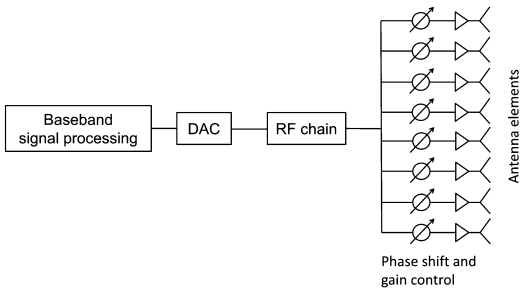

FIGURE 7.3

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.

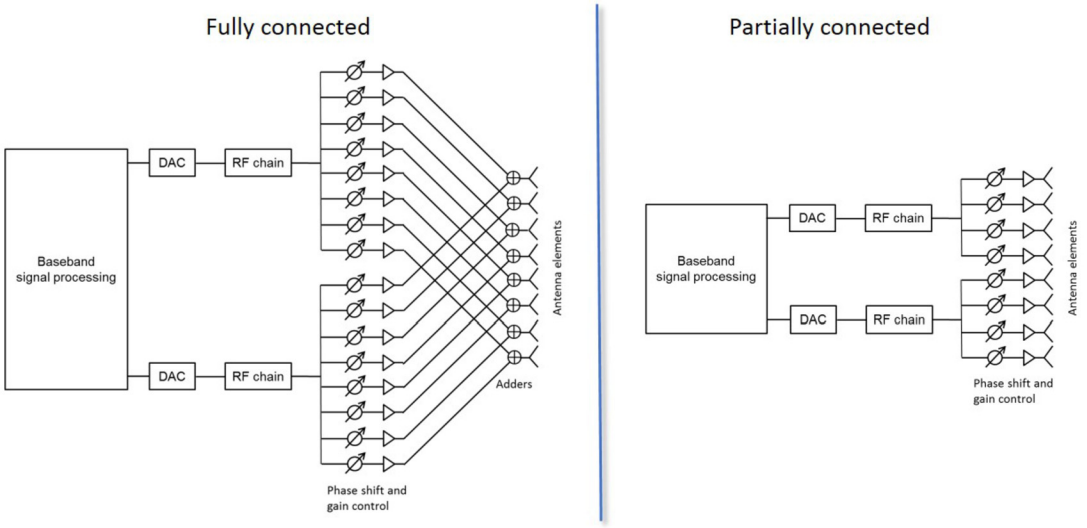

FIGURE 7.4

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.

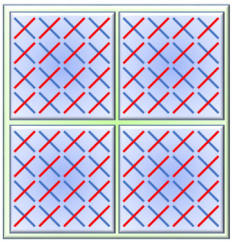

FIGURE 7.5

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.

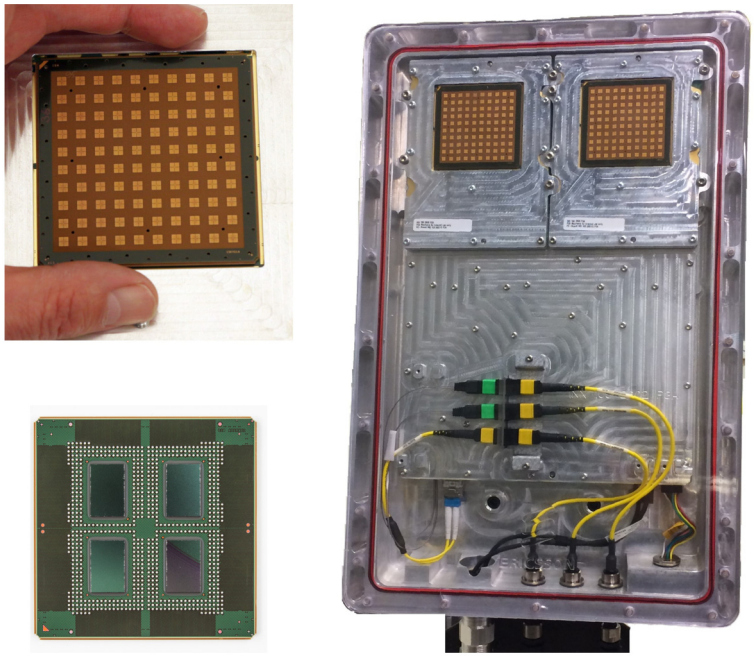

FIGURE 7.6

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

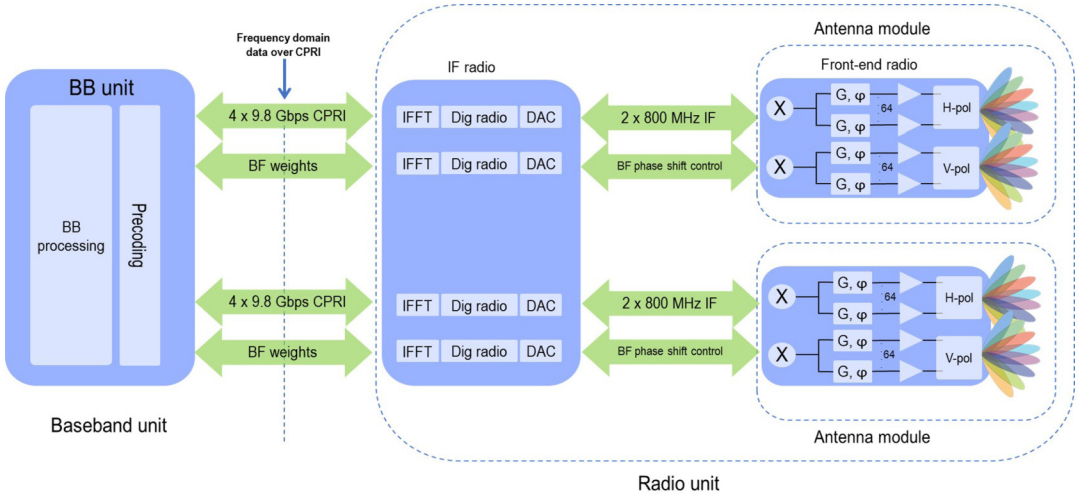

FIGURE 7.7

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

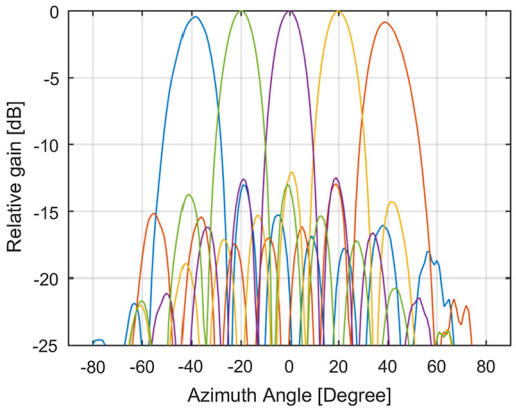

FIGURE 7.8

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.

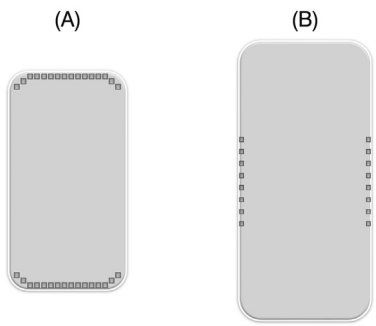

FIGURE 7.9

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.

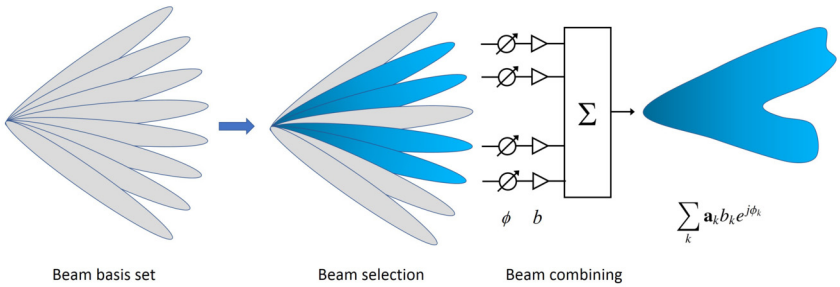

FIGURE 7.10

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

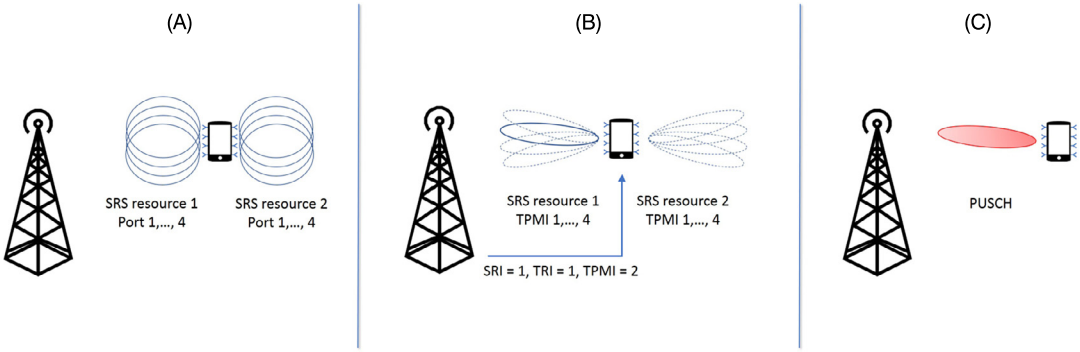

FIGURE 7.11

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.

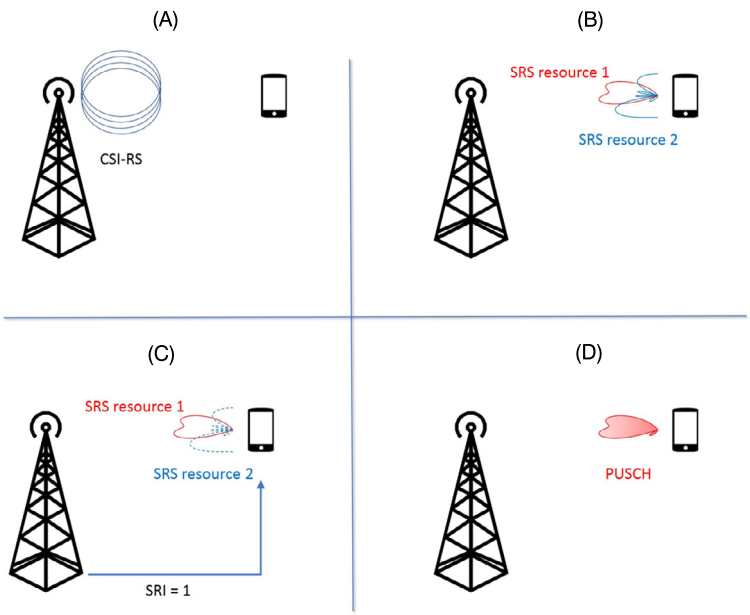

FIGURE 7.12

Non-codebook-based uplink transmission. An example of non-codebook-based uplink transmission.

Uplink Tx diversity is not explicitly specified, but open-loop Tx diversity can be performed transparently by, e.g., frequency hopping. Furthermore, Tx antenna selection diversity can be performed by utilizing SRI or CSI-RS.

7.3.4 BEAM MANAGEMENT

One of the main new features in NR is the support for analog beam-forming, which is foreseen to be prevailing at millimeter-wave frequencies. For this purpose a new framework called beam management has been developed in order to support analog beam-forming at both the BS and the UE side. Beam management has been defined in 3GPP as a set of Layer 1/2 procedures to acquire and maintain a set of BS and/or UE beams[26] that can be used for downlink and uplink transmission/reception [1]. It includes a number of features, such as:

- Sweeping. Covering an angular sector by sweeping analog beams over the sector.
- Measurement. Measuring the quality of different beams.
- Reporting. Reporting beam information such as which beams are best and their measured qualities.
- Determination. Selecting one or a few beams out of a number of candidate beams.
- Indication. Indicating which beam or beams has been or have been selected for data transmission.
- Switching. Switching to another beam if another beam gets higher quality than the current beam.
- Recovery. Finding a new beam if the current beam cannot maintain a communication link due to, e.g., blockage.

If analog beam-forming is used in both the BS and the UE, beams at both sides need to be found so that beam pairs rather than just beams need to be acquired and maintained. Beam management is used for both data and control channels, i.e., the physical downlink shared channel (PDSCH) and physical downlink control channel (PDCCH) (see Chapter 2). Multiple beam pair links can be established for robustness, joint transmission or spatial multiplexing.

Beam management is intended to be a fast process by involving only Layer 1 and 2 signaling. It is aimed at maintaining beam pair links between a UE and beams from one BS or several tightly synchronized BSs. Mobility between unsynchronized BSs is handled by other procedures involving slower Layer 3 signaling. The beam management framework has also been designed for flexibility so it can be adapted to different use cases. For example, for short data sessions, when only small amount of data should be transmitted/received, a quick beam management procedure using a coarse level of accuracy could be used, while a more refined procedure can be invoked for longer sessions.

A general principle of beam finding is to transmit reference signals in a number of candidate beams and estimate the quality of the received reference signal at the receiver for each candidate Tx beam. In NR, two different reference signal types can be used for downlink beam management: SSB and CSI-RS. For uplink beam management, SRS can be used. A reference to the beam(s) with highest quality is then reported back to the transmitter so that the transmitter knows which beam is best to use in subsequent data transmissions. For NR downlink, an index to a previously transmitted reference signal (SSB or CSI-RS) is used as a reference to the best beam, so-called SSB resource indicator

[26]The NR specifications seldom use the term "beam". What we mean by beam herein is what in the NR specifications is referred to as *spatial domain Tx filter* or *spatial domain Rx filter*.

(SSBRI) or CRI. Similarly, the receiver can find its best Rx beam by estimating the quality of the received reference signal for each candidate Rx beam. During the Rx beam finding procedure, the transmitter should repeat the transmission of the reference signal in the same Tx beam so that the receiver can make a fair comparison of the candidate Rx beams. These procedures should be repeated whenever needed to track UE movements and changes in the radio environment.

The main focus of this section is on downlink beam management. In many cases uplink beam management is not needed. If the BS and UE has so-called *beam correspondence*, the beam pair links established by downlink beam management procedures can also be used for uplink without any need for separate uplink procedures. Beam correspondence is a kind of reciprocity that implies that a BS or UE can determine its Tx beam based on measurements on its Rx beams, or vice versa. For example, if a set of beam-forming weights gives the same beam pattern for the Tx and the Rx, beam correspondence holds. If beam correspondence does not hold, separate uplink beam management procedures can be invoked. These are similar to the corresponding downlink procedures. Uplink specific beam management aspects are summarized at the end of this section.

7.3.4.1 Beam Acquisition During Initial Access

Before a UE enters the network, a BS does not have any information about the direction to the UE. The signals in the initial access procedure therefore need to be transmitted and received without any prior knowledge about the direction to the UE. At low frequencies, these signals can be transmitted with a wide beam that covers the entire angular sector of the cell. At millimeter-wave frequencies, however, the antenna gain with such a wide beam may not be sufficient to achieve the desired coverage. Therefore, NR supports beam-forming also of initial access signals. Since the direction to a UE is not known a priori, the set of beams carrying the initial access signals must cover the entire sector, e.g., by beam sweeping.

The first signals that the BS transmits to aid a UE to access the network are contained in the SSB. The purpose with SSB is to provide coarse time and frequency synchronization and basic system information such as how the UE shall access the system, indication of the physical cell identity (ID), and where to find the remaining configurations. The SSB is transmitted periodically and can be beam-formed in order to ensure sufficient cell coverage. The SSB is then transmitted repeatedly in different beams, a so-called SSB burst set. For carrier frequencies above 6 GHz, up to 64 beams can be used within an SSB burst set of 5 ms. This burst set is then repeated with a specified periodicity. For initial cell selection the default value of the SSB burst set periodicity is 20 ms. See Fig. 7.13 for an illustration of beam-forming of SSB.

During initial access, the UE measures the different SSBs in an SSB burst set in order to determine the best BS Tx beam. If the UE has analog beam-forming it can measure multiple SSB burst sets to find a suitable Rx beam, or it can use a wide beam. The UE then transmits the physical random access channel (PRACH) preamble in a resource that is associated to the SSB that corresponds to the determined BS Tx beam. If the UE has analog beam-forming and beam correspondence, the UE can transmit the preamble with a Tx beam that corresponds to the Rx beam it used when receiving the SSB from the best BS Tx beam. The BS can receive the PRACH preamble using a wide beam or the same beams it used for transmitting the SSB burst set. When the BS has received the PRACH preamble it can deduce from the corresponding PRACH resource which BS Tx beam was best for that UE. The BS can then use the identified Tx beam in possible beam refinement procedures or subsequent data and control transmission.

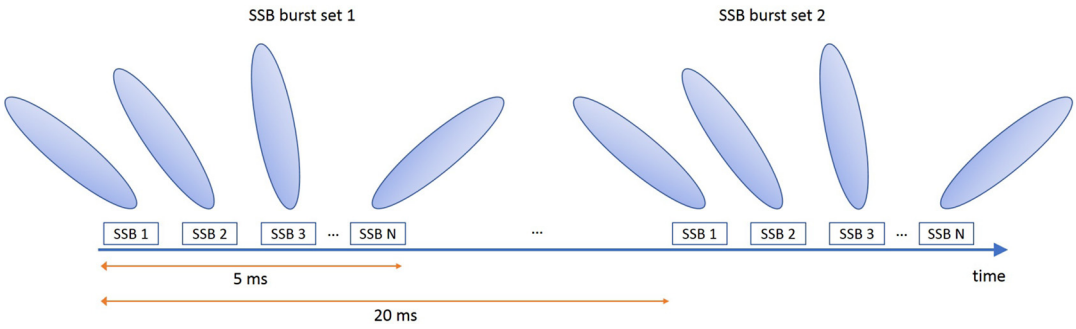

FIGURE 7.13

SSB burst. Two SSB burst sets with N beams in each set.

Since the SSB can be transmitted with a periodic beam sweep, it is also useful for beam management purposes. SSB can be used not only for initial beam acquisition but also for other purposes such as input to beam refinement procedures and discovery of new beams when a UE moves or if changes occur in the radio environment. In some cases, SSB may be all that is needed in beam management. The SSB transmissions can be used both by the BS and the UE to find suitable Tx and Rx beams.

To provide robustness against mobility and blockage and to reduce signaling overhead, the SSB beams can be relatively wide so that the cell can be covered by a few beams. For UEs with low data rate requirements or good channel conditions it may be sufficient to use the wide beams acquired during initial access also for the data transmission. In other cases, narrower beams with higher gain can be acquired by subsequent beam refinement procedures. The beams found during initial access can then be useful as input to the beam refinement procedures, e.g., to trigger a beam sweep with narrow beams centered around the direction estimated during initial access.

7.3.4.2 Beam Management Procedures

Although not explicitly stated in the specifications, downlink beam management has been divided into three procedures [1]:

- P-1. The purpose of P-1 is to find initial BS Tx beam(s) and possibly also UE Rx beam(s) by performing a beam sweep over a relatively wide angular sector.
- P-2. This is used for beam refinement of the BS Tx beam(s) by performing a beam sweep in a more narrow angular sector than in P1.
- P-3. This is used for performing an Rx beam sweep at the UE. In P-3, the BS Tx beam is fixed during the UE Rx beam sweep.

There are similarities between the procedures and not all procedures are needed. Furthermore, P-2 can be a special case of P-1. An example of how the P-1, P-2, and P-3 procedures can be performed is schematically illustrated in Fig. 7.14. In P-1, the BS performs a beam sweep over an angular sector that covers the entire cell by transmitting a unique reference signal in each beam. To limit the number of beams in such a wide beam sweep the beams could be relatively wide to give an initial, coarse estimate of the best beam direction. The reference signal could be, e.g., the SSBs during initial access

or a periodic CSI-RS transmission that has been configured for beam management. The UE measures the power of the received reference signals from all BS Tx beams using a wide Rx beam and reports to the BS which beam has the highest received power. In P-2, the BS performs beam refinement by an aperiodic CSI-RS transmission using narrower beams in an angular sector around the best beam reported by the UE in P-1. The UE measures the power of the received CSI-RSs from these BS TX beams, still using a wide Rx beam, and it reports to the BS which of the narrow beams has the highest received power. In P-3, the BS transmits CSI-RS repeatedly in the best narrow beam reported by the UE in P-2 so that the UE can perform an Rx beam sweep to find its best Rx beam by measuring the power of the received CSI-RS in each Rx beam. In the data transmission, the BS uses the best BS Tx beam found during P-2 and the UE uses the best UE Rx beam found during P-3.

Note that this is just one example of how to perform beam management and other ways are possible. For example, P-1 could be a joint BS Tx/UE Rx beam sweep in which the UE sweeps its Rx beams for each BS Tx beam. The BS then has to repeat the reference signal transmissions in each BS Tx beam so that the UE can evaluate different Rx beams for every BS Tx beam. Therefore, this approach is more costly in terms of reference signaling overhead and beam acquisition time.

To provide robustness against blocking, a UE can be configured to monitor PDCCH on multiple beam pair links. For example, while data transmission is being performed on an active beam pair link, the UE can monitor PDCCH on another beam pair as a backup link for swift fallback if there should be a sudden blockage of the active link.

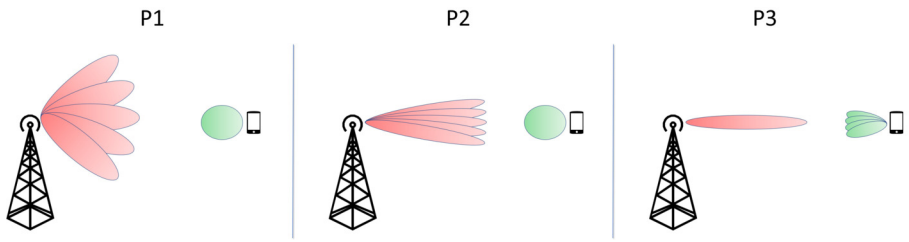

FIGURE 7.14

Beam management procedures. Schematic illustration of the beam management procedures P-1, P-2, and P-3.

Which beam management process to use is not explicitly configured. Instead, the CSI acquisition framework described previously is used to configure a UE with the pertinent report and resource settings. For beam management, a report setting can for example contain information regarding the number of beams to report on, which CSI parameters to report (e.g., L1-RSRP), time-domain behavior, and frequency granularity of the reporting. A resource setting can contain information regarding which reference signal type (CSI-RS, SSB) the UE shall measure on, time-domain behavior, and one or multiple resource sets each containing multiple CSI-RS resources.

7.3.4.3 Beam Measurement and Reporting

A central component in beam management is measuring and reporting of the quality of different candidate beams. In downlink beam management, measurements are performed by the UE on SSB or CSI-RS transmitted by the BS. Beam sweeping using CSI-RS is more flexible than SSB since there is more freedom in configuring the CSI-RS transmission. For example, the number of beams and their

coverage area can be configured more flexibly. A CSI-RS beam sweep is UE-specifically configured, so it is possible to tailor a beam sweep for a particular UE. For example, the beams in a CSI-RS beam sweep can be narrow beams centered around a wide beam acquired from SSB during initial access. A CSI-RS beam sweep does not need to be transmitted periodically but can be triggered whenever needed. Other advantages with CSI-RS over SSB for beam management are that CSI-RS is more wideband and can use two antenna ports, which potentially could lead to more reliable beam measurements. For example, two antenna ports can be used for transmitting with two orthogonal polarizations in order to reduce polarization mismatch losses. An advantage with using SSB for beam management is that SSB is transmitted for initial access purpose anyway, so using it also for beam management incurs no additional reference signal overhead.

CSI-RS for beam management can be periodic, semi-persistent, or aperiodic. Periodic and semi-persistent CSI-RS can be efficient in wide beam sweeps for acquiring coarse beams when there is little a priori information as regards the directions to UEs, or for beam sweeps at high load for covering a service area containing multiple UEs. To save signaling overhead, aperiodic CSI-RS can be more efficient at low load to perform a refined, UE-specific beam sweep when there is some prior knowledge about the direction to the UE.

The quality metric used for beam management is L1-RSRP for both SSB and CSI-RS. In the case that CSI-RS resources with two ports have been configured, the UE shall report the linear average over both ports. Information as regards which beam the UE has selected is conveyed by reporting an index to the corresponding CSI-RS resource, i.e. CRI. The UE can be configured by the network to report a number of strongest beams.

NR also supports that the UE can report a group of BS Tx beams that can be received simultaneously by the UE, e.g., by using different antenna panels. This can be useful for establishing multiple beam pair links for robustness, joint transmission or spatial multiplexing. The grouping-based reporting can be turned on and off on a per UE basis.

For non-grouping-based reporting, the UE can be configured to measure on up to 64 beams in a beam sweep and report up to four beams in an reporting instance. The reports are transmitted on physical uplink control channel (PUCCH) or PUSCH depending on the reporting being periodic, semi-persistent, or aperiodic.

Measurement and reporting for beam management and CSI acquisition have a common framework so it can be used for both purposes by different configurations. At low frequencies with digital beam-forming, beam management is typically not needed. In this case, the UE can be configured to measure on a number of CSI-RS resources, select the best resources, and report CSI for these resources. At high frequencies with hybrid beam-forming the optimal approach would be to jointly optimize the analog beam-formers and the digital precoder. However, this may not be feasible due to high computational complexity and signaling overhead. A more practical approach is to determine the analog beam-formers and digital precoder separately. The analog beam-formers can then be determined using a beam management procedure followed by design of the digital precoder using the selected analog beams and the CSI acquisition framework. For example, the BS can first transmit SSB or CSI-RS configured for beam management in order to select beam pair links. The UE measures and reports indices and qualities of the preferred beam pair links. The BS can then transmit CSI-RS for CSI acquisition using the selected beam pair links. The UE measures on the CSI-RSs and reports CQI, PMI, and RI. The BS then transmits data using the selected analog beams and digital precoder.

7.3.4.4 Beam Indication

For downlink, NR supports beam management with and without beam indication. Beam indication is an indication to the UE which Tx beam the BS will use in coming data or control transmission so that the UE can update its Rx beam accordingly. This may be needed when analog beam-forming is used in the UE. If the BS changes its Tx beam the UE may need to change its Rx beam at the same time so that it matches the new BS Tx beam, or the link may be lost. It can also be needed for the UE to know which beam to receive a CSI-RS beam sweep in a limited sector (a P-2 procedure). To be able to do this with analog beam-forming, the UE needs to know beforehand which Rx beam to use so that it can switch to that beam when the data is transmitted with the new Tx beam. The downlink beam indication provides information to the UE so that it can use an Rx beam that is suitable for the new Tx beam.

Beam indication is needed when several beam pair links are maintained, e.g., when monitoring multiple PDCCHs, or when a joint BS Tx and UE Rx beam sweep is performed. In that case all combinations of BS Tx beams and UE Rx beams are evaluated, the UE reports the best beam pairs and the BS selects which Tx beam to use. A beam indication is then needed since the UE needs to know which Tx beam the BS uses in order to set its Rx beam.

In other cases, explicit beam indication may not be needed. For example, if a single beam pair link is used for PDSCH and PDCCH transmission and the UE reports only the best BS Tx beam in a beam sweep, the UE does not need any beam indication provided that the BS uses the reported beam in the next transmission. When operating without beam indication, different BS Tx beams are evaluated under the assumption that the UE holds its Rx beam fixed and different UE Rx beams are evaluated under the assumption that the BS holds its Tx beam fixed. The UE can be informed that the BS keeps its Tx beam fixed by a CSI-RS resource set configuration that contains an information element which indicates that the BS repeats its CSI-RS transmission in the same Tx beam. For example, in a sequential P-2/P-3 procedure the BS first performs a P-2 Tx beam sweep and the UE keeps its Rx beam fixed (or uses a wide beam). After the P-2 procedure, the BS updates its Tx beam without informing the UE and triggers a P-3 UE Rx beam sweep. After the P-3 procedure, the UE updates its Rx beam without informing the BS. The UE only needs to remember which Rx beam was best and use that information in the next reception without any beam indication from the BS. An advantage with operation without beam indication is the reduced signaling overhead and delay associated with beam indication.

A beam indication is a reference to a previously transmitted reference signal in the form of a spatial QCL relation. More specifically, it is an indication that the UE can assume that the DM-RS of PDSCH/PDCCH is spatially quasi-co-located with a previously transmitted downlink reference signal, e.g., a particular SSB or CSI-RS resource. This means that the UE can use the same Rx beam for the coming data/control transmission as it used when it received the indicated reference signal. Beam indication can also be used to provide a spatial QCL relation between different reference signals, e.g., that an aperiodic CSI-RS is spatially quasi-co-located with an SSB. This could be used for, e.g., beam refinement when an aperiodic CSI-RS beam sweep should be performed with narrow beams within a wide SSB beam.

Downlink beam indication is performed by signaling a so-called transmission configuration indicator (TCI) to the UE which provides a spatial QCL reference that the UE can use to set its Rx beam. This is similar to the PDSCH rate matching and quasicolocation indicator (PQI) used for CoMP operation in LTE. The UE is configured with a list of TCI states by higher-layer signaling, where each TCI state is configured with a set of CSI-RS or SSB IDs. In each state, one CSI-RS or SSB ID that should be used as spatial QCL reference is selected. The beam indication is then performed by signaling a selected

TCI state to the UE that it should use for obtaining a QCL reference for the coming PDSCH/PDCCH transmission. By maintaining multiple TCI states, the BS can dynamically switch between different Tx beams. Before the TCI states have been configured and activated, the UE can assume that the DM-RS of PDSCH is spatially quasi-co-located with the SSB determined in the initial access procedure. This means that the UE can use the same Rx beam it used when it received the SSB during initial access. A TCI state can also contain a QCL reference for time/frequency parameters such as delay and Doppler spread. Reference signal for time/frequency QCL can be, e.g., a TRS (see Section 2.5). For frequencies below 6 GHz, a TCI state may contain only a time/frequency QCL reference and no spatial QCL information.

7.3.4.5 Beam Recovery

Narrow-beam transmission and reception is useful for improving the link budget at millimeter-wave frequencies but this may become susceptible to so-called beam failure. A beam failure means that the quality of the beam pair links for all control channels becomes too low for maintaining communication. This can be caused by, e.g., sudden blockage or failure in a beam management process. This may be lead to radio link failure (RLF) wherein a costly higher-layer reconnection procedure is needed. To avoid this, NR supports a faster procedure using lower layer signaling to recover from beam failure, referred to as beam recovery. For example, instead of initiating a cell reselection when a beam pair link quality becomes too low, a beam pair reselection within the cell can be performed.

Beam failure is detected by monitoring a beam failure detection reference signal and assessing if a beam failure trigger condition has been met. The beam failure detection reference signal can be SSB or a periodic CSI-RS configured for beam management and is associated with the Tx beam with which a control channel is transmitted. Beam failure detection is triggered if a number of consecutive detected beam failure instances exceeds a maximum value, where a beam failure instance occurs if the block error rate (BLER) of a hypothetical PDCCH transmission is above a threshold.

To find candidate new beams, the UE monitors a beam identification reference signal, which can be SSB or a periodic CSI-RS configured for beam management. These reference signals can be transmitted with wider beams than the ones used for data. They can be used both for finding a new BS Tx beam and a new UE Rx beam. When a UE has declared beam failure and found a new beam it transmits a beam recovery request message to the serving BS. The BS responds to the request by transmitting a recovery response over PDCCH to the UE and the UE monitors the control channel for the response. If the response is received successfully, the beam recovery is completed and a new beam pair link has thus been established. If the UE cannot detect any response within a specified time, the UE may perform a retransmission of the request. If the UE cannot detect any response after a specified number of retransmissions, then it notifies higher layers, potentially leading to RLF and cell reselection.

7.3.4.6 Uplink Beam Management

As mentioned previously, if the BS and UE have beam correspondence, separate procedures for uplink beam management are not needed, since the beam determination can be based on downlink procedures. If beam correspondence does not hold, separate uplink procedures similar to P-1, P-2, and P-3 described previously can be used. These are referred to as U-1, U-2, and U-3, and are based on uplink reference signals.

In uplink beam management, SRS is used in a similar way as CSI-RS is used in downlink beam management. An uplink beam management procedure is initiated by the BS by triggering one or several

SRS resource sets, where each resource set contains a number of SRS resources. One SRS resource is transmitted per UE Tx beam, the BS measures the received SRSs, determines a preferred UE Tx beam, and reports an SRI to the UE. The UE may also repeat an SRS transmission in a Tx beam so that the BS can find a suitable Rx beam. Multiple SRS resource sets can be used to train multiple beam-formers in parallel by triggering one set per beam-former.

Uplink beam indication may be needed so that the UE can adjust its Tx beam based on the Rx beam used by the BS. This may be needed also when beam correspondence holds and no separate uplink beam management is performed. Beam indication for uplink is supported for PUCCH and SRS. PUCCH can be spatially related to SRS or downlink signals such as SSB or CSI-RS. Such spatial relations can be used if the UE has beam correspondence. For example, if a UE receives a downlink RS in an Rx beam pointing in a certain direction it can perform an uplink transmission in a Tx beam pointing in the same direction. If the UE does not have beam correspondence, it can use a previously transmitted SRS for determining its Tx beam.

7.4 EXPERIMENTAL RESULTS

In order to lend some empirical support to the theoretical descriptions in previous sections, this section shows results from two different measurement campaigns that demonstrate the effectiveness of beam-forming at millimeter-wave frequencies. In the first campaign, an assessment of the achievable beam-forming gain using analog beam-forming in different environments was made and in the second campaign, successful beam tracking of a high speed UE was demonstrated. Furthermore, we present some results from system simulations using 3GPP models that illustrate what SINR levels and beam-forming gains can be expected in an urban macro scenario at 30 GHz for different antenna sizes at the BS and the UE.

7.4.1 BEAM-FORMING GAIN

In [42], measurements of beam-forming gain using analog beam-forming at 15 GHz in different environments were reported. The measurements were performed with a 5G test-bed using a planar antenna array with 8×8 antenna elements divided into 32 subarrays of 2×1 antenna elements each. Analog beam-forming using a 2-D GoB with 48 beams was applied over the subarrays to generate beams with an azimuth and elevation half-power beam width (HPBW) of 14°. Reference signals were transmitted sequentially in each beam and a UE prototype measured and logged average received power of the reference signals for all beams every 100 ms along a drive route. The analog beam grid is depicted in Fig. 7.15.

Fig. 7.16 shows estimated beam-forming gain along a drive route in an outdoor environment consisting of a square with surrounding buildings and streets. Buildings and the BS position are indicated in the sketch of the environment. The BS is mounted 7 m above the ground and faces the 110 m $\times$ 60 m square, which is surrounded by buildings of two to eight storeys. The measurement positions on the square are in LoS and the positions on the four streets out from the square are mainly in NLoS from the BS. The beam-forming gain was estimated by comparing the received power in the strongest beam with the received power averaged over all beams. The brightness of each dot on the route shows

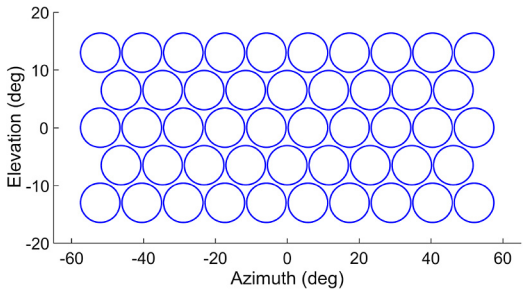

FIGURE 7.15

GoB. Beam grid used in the measurements.

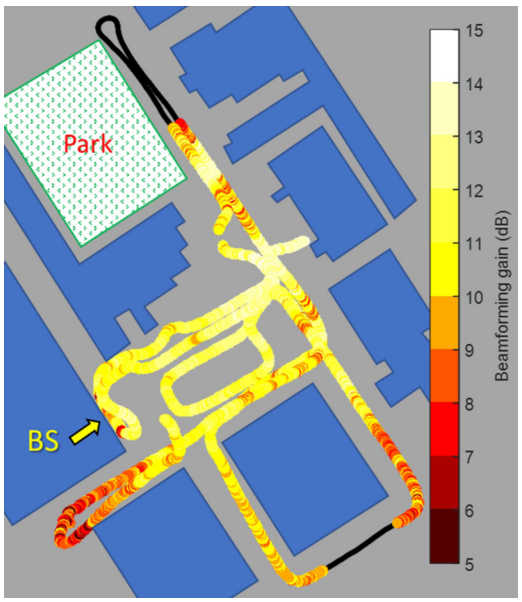

FIGURE 7.16

Beam-forming gain. Estimated beam-forming gain along an outdoor drive route.

the estimated beam-forming gain for that position.[27] Since the analog beam-forming is performed over 32 subarrays, the maximum beam-forming gain is 15 dB. It is clear from this figure that the LoS positions on the open square in front of the BS have higher beam-forming gain than the NLoS positions on the streets behind buildings. It can also be seen that the beam-forming gain is somewhat higher in

[27] The thinner black parts of the route correspond to lost connections.

the center of the square than in the positions close to the surrounding buildings where reflections have a negative impact on the beam-forming gain.

Fig. 7.17 shows cumulative distribution functions (CDFs) of the estimated beam-forming gain from the drive route shown in Fig. 7.16. The measurement results have been separated into LoS and NLoS positions along the route. The results show that the beam-forming gain is high in the LoS positions, mainly in the range 10–13 dB. As expected, the beam-forming gain in the NLoS positions is lower, around 7–12 dB. Results of measurements in an indoor environment with rich multipath were also reported in [42]. In this environment, the beam-forming gain was around 6–11 dB. Although the gain with analog beam-forming is substantial also in the reflective environments, there is potential for higher gain with digital precoding that can utilize several different propagation paths. The potential gain with hybrid beam-forming was also assessed in [42] by assuming a perfect coherent combining of the best analog beams. A substantial improvement was found by combining only a few beams. To achieve a certain total beam-forming gain, more beams were required in the reflective environments.

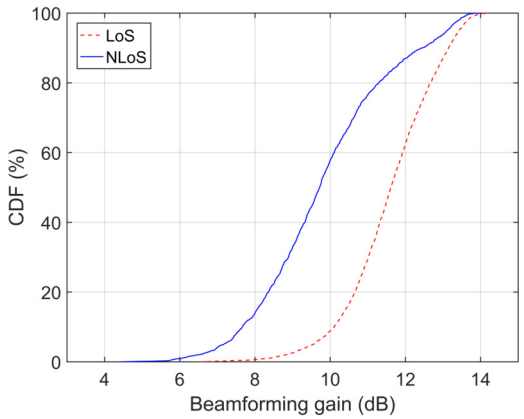

FIGURE 7.17

CDF of beam-forming gain. CDF of estimated beam-forming gain.

7.4.2 BEAM TRACKING

NR should support use cases with high mobility such as users traveling with high speed in, e.g., trains and cars. At millimeter-wave frequencies it may be essential to track moving users with narrow beams using analog beam-forming. Experimental results of beam tracking at 28 GHz of a car moving at high speed have been reported in [26]. The measurements were performed with a test system consisting of four transmission points deployed along a race track in Republic of Korea and one UE mounted on top of a car. Each transmission point had a rectangular antenna array with $16 \times 4 = 64$ antenna elements per polarization. Analog beam-forming was performed by using a GoB shown in Fig. 7.18. The HPBW of the array was 6° in azimuth and 24° in elevation. The UE had four dual-polarized, directive antennas covering different angular sectors. The bandwidth of the system was 800 MHz and supported MIMO transmission of two layers providing a peak rate of about 7.8 Gbps.

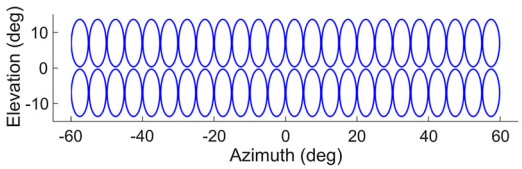

FIGURE 7.18

GoB. Beam grid used in the measurements.

Beam tracking was performed by transmitting beam-formed reference signals which were measured upon and reported by the UE. In the beam selection, beams from all four transmission points were evaluated and the best beam was selected regardless from which antenna the beam was transmitted. The transmission points were connected to the same baseband. Drive tests were performed on the race track at speeds up to 170 km/h. Fig. 7.19 shows the downlink throughput and UE speed as a function of time along the drive route. It can be seen that at low speed the throughput exceeded 6 Gbps and at the highest speed of 170 km/h, a throughput of up to 3.6 Gbps was achieved. Furthermore, successful beam switching was reported in [26], even though the car stayed in a beam for only 25 ms at high speeds.

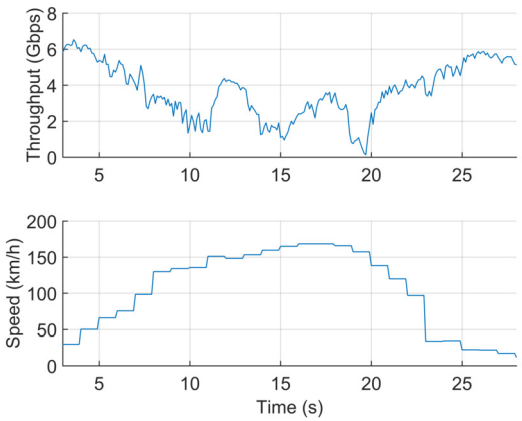

FIGURE 7.19

Throughput. Downlink throughput and UE speed vs. time.

7.4.3 SYSTEM SIMULATIONS

Finally, this section provides some results from system simulations using a 3GPP NR scenario and channel model that show how analog and hybrid beam-forming can increase the SINR at millimeter-wave frequencies.

Ideally, the analog beam-forming gain is $10\log_{10}(N_T N_R)$ (dB), where N_T and N_R is the number of Tx and Rx antenna elements, respectively. However, as discussed in Section 7.2.1, this can only be achieved in highly correlated channels when a LoS or specular reflection direction is known both at the transmitter and the receiver so that a beam can be steered in this direction. In practice angular spread will limit the analog beam-forming gain. Furthermore, with analog beam-forming a limited number of hypothesized directions must be tested sequentially in time to find the best beam direction, typically by using a fixed GoB. Since the true direction will not coincide exactly with an angle grid point, a so-called straddling loss will incur. To evaluate the beam-forming gains that are more realistic to achieve, we show results from system simulations based on models that take these effects into account.

We evaluate the downlink SINR as a function of the number of antenna elements in a BS panel for an urban macro scenario at 30 GHz carrier frequency using the 5G channel model[28] in [2]. The inter-site distance is 200 m and other simulation assumptions are as in [37]. We assume that the BS has one, two, or four square panels with $N_T \times N_T$ dual-polarized antenna elements each. Furthermore, we assume that analog beam-forming is performed per polarization within each panel and that the panels and polarizations are combined to a single transmit layer using SVD precoding assuming perfect CSIT. Two different UE antenna configurations are considered: one with and one without beam-forming. For the case without beam-forming the UE is assumed to have one dual-polarized isotropic antenna and for the case with beam-forming the UE is assumed to have two panels mounted back-to-back, where each panel has 2×4 dual-polarized antenna elements. During data reception only the UE panel with highest received power during the beam management procedure is used, i.e., panel combining is not used in the UE. It assumed that analog beam-forming is performed per polarization within this panel and that the polarizations are combined with interference rejection combining.

An ideal beam management procedure is simulated by finding the best pair of analog Tx beam at the BS and analog Rx beam at the UE. The beams are selected from a GoB constructed from DFT beams without angular oversampling. The zenith angles in the BS GoB are given by the angles $180^\circ \cdot (n - 1/2)/N, n = 1, \ldots, N$, where N is the number of antenna rows in a panel, that lie in the interval [90° 160°]. The azimuth angles in the BS GoB are given by the angles $180^\circ \cdot (n - 1/2)/N - 90^\circ$, $n = 1, \ldots, N$, where N is the number of antenna columns in a panel that lie in the interval $[-60^\circ\ 60^\circ]$. The same analog beam is used in both polarizations and in all BS panels in the case that it has multiple panels.

Fig. 7.20 shows the 5- and 50-percentile, respectively, in the SINR CDF over all UEs in the network as a function of the number of BS antenna elements, N_T^2, per panel assuming square panels each having $N_T \times N_T$ elements. The different curves are for different numbers of BS and UE panels. The solid curves are for the cases with one, two, or four BS panels and two UE panels. The dashed curves are the corresponding results for a single dual-polarized antenna at the UE.

The results show that without any BS or UE beam-forming the performance is not acceptable, since the 5-percentile SINR is below −16 dB. With a single 8×8 BS panel and two 2×4 panels in the UE, the 5-percentile SINR is increased to −4 dB, which is a significant improvement. The results also show that doubling the number of BS panels gives roughly a 3 dB increase in SINR. Doubling the number of antenna elements within a panel gives slightly less than 3 dB 5-percentile SINR gain for

[28] However, recall from Chapter 3 that the highly resolved properties of the channel model have not been experimentally validated.

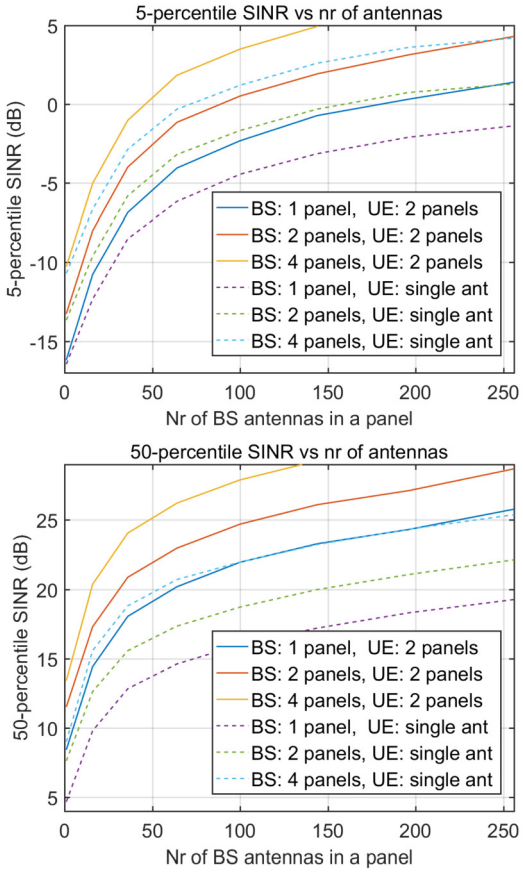

FIGURE 7.20

Simulation results. SINR vs. the number of antenna elements in a BS antenna panel obtained from system simulations using the 3GPP urban macro 5G channel model.

large panels. This may be due to the angular spread in the channel combined with narrow analog beams. Note that the multiple panels in the BS have in these simulations been used for coherent combining to increase coverage. Alternatively, the panels could be used to increase capacity by, e.g., MU-MIMO transmission.

The SINR gain of beam-forming in the UE, i.e., going from a single dual-polarized antenna in the UE to switching between two 2×4 panels, is around 2–3 dB at the 5-percentile and around 5–6 dB at the median. A possible explanation of the lower SINR gain at the 5-percentile is that UEs with low SINR are probably located indoors and/or in NLoS and therefore experience a larger angular spread which will limit the analog beam-forming gain.

REFERENCES

[1] 3GPP TR 38.802, Study on New Radio Access Technology, Physical Layer Aspects, 2017.
[2] 3GPP TR 38.900, Study on channel model for frequency spectrum above 6 GHz, 2017.
[3] 3GPP TR 38.901, Study on channel model for frequencies from 0.5 to 100 GHz, 2017.
[4] 3GPP TS 36.211, Physical channels and modulation, 2018.
[5] 3GPP TS 38.101-1, NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone, 2017.
[6] S.M. Alamouti, A simple transmit diversity technique for wireless communications, IEEE Journal on Selected Areas in Communications 16 (8) (1998, Oct.) 1451–1458.
[7] X. An, C.S. Sum, R.V. Prasad, J. Wang, Z. Lan, J. Wang, R. Hekmat, H. Harada, I. Niemegeers, Beam switching support to resolve link-blockage problem in 60 GHz WPANs, in: 2009 IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications, 2009, Sept., pp. 390–394.
[8] J.B. Andersen, Array gain and capacity for known random channels with multiple element arrays at both ends, IEEE Journal on Selected Areas in Communications 18 (11) (2000, Nov.) 2172–2178.
[9] E. Björnson, M. Bengtsson, B. Ottersten, Optimal multiuser transmit beamforming: a difficult problem with a simple solution structure, IEEE Signal Processing Magazine 31 (4) (2014, July) 142–148.
[10] G. Caire, S. Shamai, On the achievable throughput of a multiantenna Gaussian broadcast channel, IEEE Transactions on Information Theory 49 (7) (2003, July) 1691–1706.
[11] D. Colombi, B. Thors, C. Törnevik, Implications of EMF exposure limits on output power levels for 5G devices above 6 GHz, IEEE Antennas and Wireless Propagation Letters 14 (2015) 1247–1249.
[12] M. Costa, Writing on dirty paper, IEEE Transactions on Information Theory 29 (3) (1983, May) 439–441.
[13] E. Dahlman, S. Parkvall, J. Sköld, 4G LTE-Advanced Pro and The Road to 5G, third ed., Academic Press, 2016.
[14] R.F.H. Fischer, C. Windpassinger, A. Lampe, J.B. Huber, Space–time transmission using Tomlinson–Harashima precoding, in: Proc. 4th Int. ITG Conf. Source and Channel Coding, 2002, pp. 139–147.
[15] X. Gu, D. Liu, C. Baks, O. Tageman, B. Sadhu, J. Hallin, L. Rexberg, A. Valdes-Garcia, A multilayer organic package with 64 dual-polarized antennas for 28 GHz 5G communication, in: 2017 IEEE MTT-S International Microwave Symposium (IMS), 2017, June, pp. 1899–1901.
[16] H2020-ICT-671650-mmMAGIC/D5.2, mmMAGIC Deliverable D5.2, Final multinode and multiantenna transmitter and receiver architectures and schemes, 2017.
[17] P. Harris, W.B. Hasan, H. Brice, B. Chitambira, M. Beach, E. Mellios, A. Nix, S. Armour, A. Doufexi, An overview of massive MIMO research at the University of Bristol, in: Radio Propagation and Technologies for 5G (2016), 2016, Oct., pp. 1–5.
[18] B.M. Hochwald, C.B. Peel, A.L. Swindlehurst, A vector-perturbation technique for near-capacity multiantenna multiuser communication-part II: perturbation, IEEE Transactions on Communications 53 (3) (2005, March) 537–544.
[19] W. Hong, K.H. Baek, Y. Lee, Y. Kim, S.T. Ko, Study and prototyping of practically large-scale mmWave antenna systems for 5G cellular devices, IEEE Communications Magazine 52 (9) (2014, September) 63–69.
[20] A. Huebner, F. Schuehlein, M. Bossert, E. Costa, H. Haas, A simple space-frequency coding scheme with cyclic delay diversity for OFDM, in: 2003 5th European Personal Mobile Communications Conference (Conf. Publ. No. 492), 2003, April, pp. 106–110.
[21] IEEE Std 145-1993, IEEE Standard Definitions of Terms for Antennas, 1993.
[22] ITU-R M.2135-1, Guidelines for evaluation of radio interface technologies for IMT-Advanced, 2009.
[23] S.A. Jafar, A. Goldsmith, Transmitter optimization and optimality of beamforming for multiple antenna systems, IEEE Transactions on Wireless Communications 3 (4) (2004, July) 1165–1175.
[24] M. Joham, W. Utschick, J.A. Nossek, Linear transmit processing in MIMO communications systems, IEEE Transactions on Signal Processing 53 (8) (2005, Aug.) 2700–2712.
[25] J. Jose, A. Ashikhmin, T.L. Marzetta, S. Vishwanath, Pilot contamination and precoding in multicell TDD systems, IEEE Transactions on Wireless Communications 10 (8) (2011, August) 2640–2651.
[26] K. Larsson, B. Halvarsson, D. Singh, R. Chana, J. Manssour, M. Na, C. Choi, S. Jo, High-speed beam tracking demonstrated using a 28 GHz 5G trial system, in: VTC 2017 Fall, 2017.
[27] X. Li, E. Björnson, E.G. Larsson, S. Zhou, J. Wang, A multicell MMSE precoder for massive MIMO systems and new large system analysis, in: 2015 IEEE Global Communications Conference (GLOBECOM), 2015, Dec., pp. 1–6.
[28] X. Li, E. Björnson, E.G. Larsson, S. Zhou, J. Wang, Massive MIMO with multicell MMSE processing: exploiting all pilots for interference suppression, EURASIP Journal on Wireless Communications and Networking (2017, June), https://doi.org/10.1186/s13638-017-0879-2.

[29] D.J. Love, R.W. Heath, V.K.N. Lau, D. Gesbert, B.D. Rao, M. Andrews, An overview of limited feedback in wireless communication systems, IEEE Journal on Selected Areas in Communications 26 (8) (2008, October) 1341–1365.
[30] R.J. Mailloux, Phased Array Antenna Handbook, second ed., Artech House, 2005.
[31] T. Marzetta, Noncooperative cellular wireless with unlimited numbers of base station antennas, IEEE Transactions on Wireless Communications 9 (11) (2010, November) 3590–3600.
[32] A. Paulraj, R. Nabar, D. Gore, Introduction to Space–Time Wireless Communications, Cambridge University Press, ISBN 9780521826150, 2003.
[33] C.B. Peel, B.M. Hochwald, A.L. Swindlehurst, A vector-perturbation technique for near-capacity multiantenna multiuser communication-part I: channel inversion and regularization, IEEE Transactions on Communications 53 (1) (2005, Jan.) 195–202.
[34] 5G collaboration in lead up to 2018 winter sporting events, press release, available at https://www.ericsson.com/en/news/2017/5/5g-collaboration-in-lead-up-to-2018-winter-games, 2017, May.
[35] IBM & Ericsson announce research advance for 5G communications networks, press release, available at https://www.ericsson.com/en/press-releases/2017/2/ibm--ericsson-announce-research-advance-for-5g-communications-networks, 2017, Feb.
[36] Going Massive with MIMO, available at https://www.ericsson.com/en/news/2018/1/massive-mimo-highlights, 2018, Jan.
[37] R1-1703536, Evaluation assumptions for Phase 2 NR MIMO system level calibration, 3GPP TSG RAN WG1 Meeting #88, 2017, February.
[38] R1-1708688, Codebook design for Type II CSI feedback, 3GPP TSG RAN WG1 Meeting #89, 2017, May.
[39] M. Sadek, A. Tarighat, A.H. Sayed, A leakage-based precoding scheme for downlink multiuser MIMO channels, IEEE Transactions on Wireless Communications 6 (5) (2007, May) 1711–1721.
[40] B. Sadhu, Y. Tousi, J. Hallin, S. Sahl, S. Reynolds, Ö. Renström, K. Sjögren, O. Haapalahti, N. Mazor, B. Bokinge, G. Weibull, H. Bengtsson, A. Carlinger, E. Westesson, J.E. Thillberg, L. Rexberg, M. Yeck, X. Gu, D. Friedman, A. Valdes-Garcia, A 28 GHz 32-element phased-array transceiver IC with concurrent dual polarized beams and 1.4 degree beam-steering resolution for 5G communication, in: 2017 IEEE International Solid-State Circuits Conference (ISSCC), 2017, Feb., pp. 128–129.
[41] H. Sampath, Linear precoding and decoding for multiple input multiple output (MIMO) wireless channels, Ph.D. thesis, Stanford University, 2001.
[42] A. Simonsson, M. Thurfjell, B. Halvarsson, J. Furuskog, S. Wallin, S. Itoh, H. Murai, D. Kurita, K. Tateishi, A. Harada, Y. Kishiyama, Beamforming gain measured on a 5G test-bed, in: 2017 IEEE 85th Vehicular Technology Conference (VTC Spring), 2017, June, pp. 1–5.
[43] V. Tarokh, H. Jafarkhani, A.R. Calderbank, Space–time block codes from orthogonal designs, IEEE Transactions on Information Theory 45 (5) (1999, Jul.) 1456–1467.
[44] S.H. Tsai, Transmit equal gain precoding in Rayleigh fading channels, IEEE Transactions on Signal Processing 57 (9) (2009, Sept.) 3717–3721.
[45] S.H. Tsai, An equal gain transmission in MIMO wireless communications, in: 2010 IEEE Global Telecommunications Conference GLOBECOM 2010, 2010, Dec., pp. 1–5.
[46] D. Tse, P. Viswanath, Fundamentals of Wireless Communications, 2005.
[47] A.M. Tulino, A. Lozano, S. Verdu, Capacity-achieving input covariance for single-user multiantenna channels, IEEE Transactions on Wireless Communications 5 (3) (2006, March) 662–671.
[48] D. Tuninetti, On the capacity of the AWGN MIMO channel under per-antenna power constraints, in: 2014 IEEE International Conference on Communications (ICC), 2014, June, pp. 2153–2157.
[49] M.K. Varanasi, T. Guess, Optimum decision feedback multiuser equalization with successive decoding achieves the total capacity of the Gaussian multiple-access channel, in: Conference Record of the Thirty-First Asilomar Conference on Signals, Systems and Computers (Cat. No.97CB36136), vol. 2, 1997, Nov., pp. 1405–1409.
[50] J. Vieira, F. Rusek, O. Edfors, S. Malkowsky, L. Liu, F. Tufvesson, Reciprocity calibration for massive MIMO: proposal, modeling, and validation, IEEE Transactions on Wireless Communications (ISSN 1536-1276) 16 (5) (2017, May) 3042–3056.
[51] H. Weingarten, Y. Steinberg, S.S. Shamai, The capacity region of the Gaussian multiple-input multiple-output broadcast channel, IEEE Transactions on Information Theory 52 (9) (2006, Sept.) 3936–3964.
[52] B. Yu, K. Yang, C.Y.D. Sim, G. Yang, A novel 28 GHz beam steering array for 5G mobile device with metallic casing application, IEEE Transactions on Antennas and Propagation 66 (1) (2018, Jan.) 462–466.
[53] X. Zheng, Y. Xie, J. Li, P. Stoica, MIMO transmit beamforming under uniform elemental power constraint, in: 2007 IEEE 8th Workshop on Signal Processing Advances in Wireless Communications, 2007, June, pp. 1–5.