CHAPTER 4

New Radio Access Physical Layer Aspects (Part 2)

In this chapter, we discuss the theoretical and practical aspects of the downlink and uplink physical layer signal processing in NR and highlight the functional and procedural similarities and differences with the LTE physical layer processing. The chapter will describe generation, configuration, and beamformed transmission of various physical signals and physical channels as well as the HARQ protocols and power control schemes. Unlike LTE where the downlink and uplink waveforms and multiple access schemes are different, the NR uses OFDM waveform as the basis for both downlink and uplink transmission (except in certain cases where DFT precoding is used in the uplink), resulting in many similarities in functional blocks and their respective operation in the downlink and uplink. The physical channel processing in NR utilizes polar codes for robust coding of the control channels and low-density parity check (LDPC) codes for the data channels, deviating from channel coding schemes that are used in LTE.

Massive MIMO is one of the main enabling technologies in 5G wireless communications. A large number of antenna elements at the base station bring extra degrees of freedom for increasing the throughput and considerable beamforming gains for improving the coverage. In practice, a large number of antenna elements can be assembled into multiple antenna panels for the purpose of cost reduction and power saving. Multi-panel MIMO is expected to become promising for mmWave massive MIMO systems. The NR enables multi-panel antenna array operation through introduction of new reference signals, measurement, and reporting procedures. In this chapter, two of the unique NR MIMO features, that is, modular and high-resolution channel state information acquisition and beam management that distinguish NR from LTE, are described. The modular framework is composed of three components, namely resource setting, CSI reporting setting, and measurement setting, which associates a resource setting with a reporting setting. These settings serve as building blocks that allow the network to customize the CSI measurement and reporting for a UE. To improve user throughput, a high-resolution dual-stage precoding referred to as Type II CSI is supported to allow more accurate estimation of the channel, thereby improving the efficiency of the NR MU-MIMO schemes. The associated codebook features a frequency non-selective basis subset selection coupled with a frequency-selective linear combination of amplitude and phase of the precoding vectors within the basis subset. As a result, the

5G NR. DOI: https://doi.org/10.1016/B978-0-08-102267-2.00021-X

precoding matrix indicator (PMI) for Type II CSI consists of several components, each with different frequency resolution. Since the NR is primarily geared toward MU-MIMO operation, Type II CSI is complemented by Type I CSI designed for scenarios that do not require high spatial resolution, for example, SU-MIMO transmission.

In order to establish and sustain a link for data transmission and reception, beam management enables the network to perform beam switching using physical-layer measurement and link quality reporting. Beam management is especially relevant for above-6 GHz frequency planning where both gNB and UE employ narrow beams for data transmission and reception. The beam management can further be used for sub-6 GHz multi-TRP scenarios. When used in conjunction with CSI acquisition, the beam management allows the network to establish a seamless and low-latency link with the UE for data transmission. This is specifically important for over-6 GHz where a large number of narrow analog beams are used for data transmission, which in some scenarios requires frequent beam switching. Once the link is established via beam management, CSI acquisition can assist the network in link adaptation.

4.1 Downlink Physical Layer Functions and Procedures

4.1.1 Overall Description of Downlink Physical Layer

The NR downlink physical layer consists of higher layer configurable functional blocks and protocols that are configured according to the downlink physical channel characteristics, use case, deployment scenario, etc. As shown in Fig. 4.1, the downlink physical layer processing generally includes receiving higher layer data [e.g., MAC PDUs in the case of downlink shared channel or master information block (MIB) in the case of physical broadcast channel (PBCH)]; cyclic redundancy check (CRC) calculation and attachment; channel encoding and rate matching; modulation; mapping to physical resources and antennas; multi-antenna processing; and support of layer-1 control and HARQ-related signaling.

It was mentioned in Chapter 3 that OFDM was chosen as the default waveform in NR for both downlink and uplink directions due to its robustness to multipath delay spread and frequency-selectivity of wireless channels as well as scheduling flexibility for transmission of different channels and signals. Unlike LTE, the DFT-precoded OFDM is an optional transmission scheme in NR uplink that is used in link-budget–limited use cases. While the use of DFT-precoded OFDM in the uplink has certain advantages in reducing the PAPR (and alternatively the cubic metric) and achieving higher power–amplifier efficiency, it has several drawbacks including limitation in the use of spatial multiplexing, asymmetric downlink/uplink transmissions which would limit the sidelink operation, and scheduling complexity.

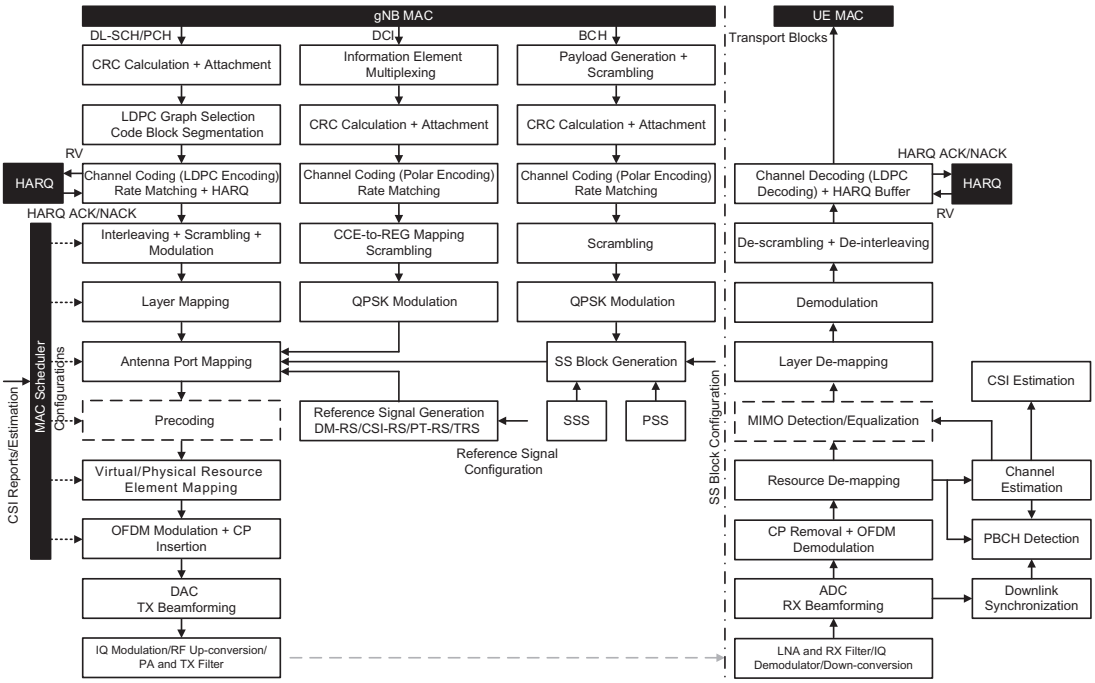


Figure 4.1
Overall downlink physical layer processing [5].

4.1.2 Reference Signals

To facilitate estimation of the multipath communication channel and reliable, coherent detection of traffic/control channels, an OFDM system makes use of reference signals (or pilot subcarriers). The pilot subcarriers provide estimate of the channel frequency response at the pilot locations over the time-frequency resource grid. It is possible to estimate the channel at other time-frequency locations using interpolation techniques. Using predefined pilot subcarriers to estimate the channel matrix, it is possible to equalize the effects of the channel and to reduce noise and interference effects on the received resource blocks. The NR specifications include several types of reference signals that are configured and transmitted in different manners, which are used for different purposes by a receiving device.

While perfect knowledge of the radio channel can be used to find an upper bound for system performance, such knowledge is not available in practice and the channel needs to be frequently estimated. Channel estimation can be performed in various ways including the use of frequency and/or time correlation properties of the wireless channel, blind or pilot-based channel estimation, and adaptive or non-adaptive channel estimation. Non-parametric methods attempt to estimate the frequency response without relying on a

specific channel model. In contrast, the parametric estimation methods assume a certain channel model and determine the parameters of this model. Spaced-time and spaced-frequency correlation functions, discussed in the previous chapter, are specific properties of channel that can be incorporated in the estimation method, improving the quality of estimations. Pilot-based estimation methods are the most commonly used in OFDM systems which are applicable in systems where the sender transmits some known signals to the receiver.

A pilot-based channel estimation is defined as the use of the channel samples estimated at the pilot tones to reconstruct channel samples at the remaining data/control-bearing subcarriers. As a result, the pilot pattern design is essentially a conventional sampling rate selection problem in the two-dimensional signal processing space. To avoid aliasing during reconstruction of the channel time-frequency function, the pilot tone selection should follow the two-dimensional sampling theorem. When multiple antennas are used, the receiver must estimate the channel impulse response (or the transfer function) from each of the transmit antennas to correctly detect the signal. This is achieved through distributing reference signals (or pilot tones) among the transmit antennas. Let $\Delta f = 1/T_u$ and $T_{symbol} = T_g + T_u$ denote the subcarrier spacing (SCS) and the OFDM symbol duration (inclusive of the guard interval), respectively. Let us further assume that the pilot subcarriers are transmitted at integer multiples of subcarrier spacing and OFDM symbol duration in frequency and time directions, respectively (i.e., $f_p = m/T_u$ and $T_p = nT_{symbol}$ where m and n are integers). The (m, n) pair represents the pilots' separation in terms of subcarrier spacing and OFDM symbol duration. From the sampling theorem point of view, the channel's two-dimensional delay-Doppler response $h(\tau, \nu)$ can be fully reconstructed, if the two-dimensional transform function $H(t,f)$ is sampled greater than or equal to the Nyquist rate across time and frequency dimensions. Hence, the time-domain sampling rate must be greater than or equal to the channel's maximum Doppler spread, that is, $T_p \leq 1/\nu_{\max}$ (the sampling rate in time must be less than the coherence time) and the frequency-domain sampling rate must be greater than or equal to the channel's maximum delay spread, that is, $f_p \leq 1/\tau_{\max}$ (the sampling rate in frequency must be less than coherence bandwidth). Assuming a wide-sense stationary uncorrelated scattering channel model and further assuming the channel to be constant over one OFDM symbol, the frequency response $H(t,f)$ of an L-path channel is given $H(t,f) = \sum_{l=1}^{L} \exp[j(\psi_l + 2\pi\nu_l t - 2\pi f\tau_l)]/\sqrt{L}$ where ψ_l, ν_l, and τ_l denote the phase, Doppler frequency, and delay of the lth path, respectively. All these parameters are independent random variables. In general, the pilot signals are oversampled to ensure a good trade-off between performance and overhead. Therefore the choice of (m, n) depends on the channel's maximum delay spread and maximum Doppler spread and must satisfy the following equation according to two-dimensional sampling theorem:

$$n \leq \frac{1}{2T_{symbol}\nu_{\max}}, \quad m \leq \frac{T_u}{2\tau_{\max}}$$

It should be noted that the pilot density for a regular pattern can be calculated using the preceding equation. For large values of $\nu_{\max}$ (i.e., large Doppler spread, or alternatively small channel coherence time means that channel is time-varying), n should be small to appropriately track channel time variations. On the other hand, for large values of $\tau_{\max}$ (i.e., large delay spread, or alternatively small coherence bandwidth means that channel is frequency-selective), m should be small to closely follow channel frequency variation. In a regularly spaced pilot pattern, the pilot symbols are evenly spaced in frequency and in time.

Cell-specific reference signals (CRS) were originally defined in 3GPP LTE Rel-8 and have continued to be used as downlink wideband always-on power-boosted pilot subcarriers which are scaled with the number of transmit antenna ports that are essential in coherent decoding/detection of the LTE downlink control channels, mobility measurements, etc. The always-on and power-boosted properties of CRS have been the potential cause of inter-cell interference in LTE networks even in the absence of user traffic. The reference signals in NR are different from LTE in the sense that they are only present when the UE has data allocation; thus they are UE specific and confined to the time-frequency region where user data are allocated. Unlike LTE, NR does not utilize cell-specific reference signals, rather exclusively relies on user-specific demodulation reference signals (DM-RSs) for channel estimation, enabling efficient beamforming and other multi-antenna schemes. In NR, the reference signals are not transmitted unless there are data to transmit, thereby improving network energy efficiency and reducing inter-cell interference. Support for low-latency transmission is an important part of NR design. In a front-loaded DM-RS structure, the reference signals and downlink control channel carrying scheduling information are located at the beginning of the slot; thus a device can start processing the received data immediately without prior buffering and time-domain interleaving across OFDM symbols, thereby minimizing the decoding delay. There are other types of reference signals such as phase tracking reference signals (PT-RS) which are used to counter phase noise at higher frequencies. A question may arise that if there are no wideband cell-specific reference signals, how the UEs will measure the reference signal received power (RSRP) during initial access or cell selection (mobility measurements) as they will not have any allocation at that time? The answer lies in the fact that NR uses a primary synchronization signal (PSS)/secondary synchronization signal (SSS)/PBCH block structure, where the PBCH has its own set of reference signals which will always be present in the PBCH so the UE while detecting the PSS/SSS/PBCH block should be able to measure an RSRP value from the PBCH DM-RS.

To support channel tracking, different types of reference signals are transmitted in downlink and uplink. The reference signals in the downlink include the following [6]:

- UE-specific DM-RS for physical downlink control channel (PDCCH) can be used for downlink channel estimation and coherent demodulation of PDCCH. The DM-RS for PDCCH is transmitted together with the PDCCH and is present only in the resource blocks that are used for PDCCH transmission.

- UE-specific DM-RS for physical downlink shared channel (PDSCH) can be used for downlink channel estimation for coherent demodulation of PDSCH. The DM-RS for PDSCH is transmitted together with the PDSCH and is present only in the resource blocks that are allocated for PDSCH transmission.
- UE-specific PT-RS can be used in addition to the DM-RS for PDSCH for correcting common phase error (CPE) between PDSCH symbols not containing DM-RS. It may also be used for Doppler and time-varying channel tracking. The phase noise of the transmitters increases as the frequency of operation increases. The PT-RS plays a crucial role especially in mmWave frequencies to minimize the effect of the oscillator phase noise on system performance. The phase noise appears as a common phase rotation of all the subcarriers, known as CPE in an OFDM system. The NR system typically maps the PT-RS information to a few subcarriers per symbol because the phase rotation equally affects all subcarriers over an OFDM symbol but exhibits low correlation from symbol to symbol. The system configures the PT-RS depending on the quality of the oscillators, carrier frequency, SCS, and modulation and coding schemes that are used for the transmission. The PT-RS for PDSCH is transmitted together with the PDSCH on need basis. The PT-RS is denser in time domain but sparser in frequency domain compared to the DM-RS, and if configured, occurs only in conjunction with the DM-RS.
- UE-specific CSI-RS can be used for estimation of CSI to allow CSI measurement and reporting which assists the gNB in modulation coding scheme (MCS) selection, resource allocation, beamforming, and MIMO rank selection. The CSI-RS can be configured for periodic, aperiodic, or semi-persistent transmission with a configurable density by the gNB. The CSI-RS also can be used for interference measurement (IM) and fine frequency/time tracking purposes. Specific instances of CSI-RS can be configured for time/frequency tracking and mobility measurements. In the absence of the cell-specific reference signals in NR the CSI-RS can be used for radio resource management, measurements and mobility management purposes in connected mode.
- Tracking reference signals (TRS) are sparse set of reference signals, which are intended to assist the device in time and frequency tracking. The TRS does not exist independently, and a specific CSI-RS configuration is used as TRS. In addition to time and frequency tracking, the TRS is used for estimation of delay spread and Doppler spread at the UE side. It is transmitted with a limited bandwidth for a configurable period of time, controlled by the upper layer parameters.

Table 4.1 provides the L1 overhead associated with various NR downlink reference signals. In NR, the overhead due to the L1/L2 control signaling depends on the size and periodicity of the configured control resource set (CORESET) in the cell which includes the overhead from the PDCCH DM-RSs. If the CORESET is transmitted in every slot, maximum control channel overhead is 21% assuming three symbols and the entire carrier bandwidth used for CORESET, while a more typical overhead is 7% when one-third of the time and frequency

Table 4.1: Various downlink reference signals and their corresponding overhead [73,74].

Reference Signal Type	Description	Overhead
PDSCH DM-RS	The DM-RS can occupy 1/3, 1/2 or one full OFDM symbol. 1, 2, 3 or 4 symbols per slot can be configured to carry DM-RS.	2.4%–29%
PDSCH PT-RS	One resource element in frequency domain every second or fourth resource block. PT-RS is mainly intended for FR2.	0.2%-0.5%
CSI-RS	One resource element per resource block per antenna port per CSI-RS periodicity	0.25% for eight antenna ports transmitted every 20 ms with 15 kHz subcarrier spacing
TRS	Two slots with two symbols in each with comb-4 configuration	0.36% or 0.18% for 20 ms and 40 ms periodicity, respectively and 15 kHz subcarrier spacing

resources in the first three symbols of a slot are allocated to PDCCH. The overhead due to the SS/PBCH block is given by the number of SS/PBCH blocks transmitted within the SS/PBCH block period, the SS/PBCH block periodicity and the subcarrier spacing. Assuming 100 resource blocks across the carrier, the overhead for 20 ms periodicity is in the range of 0.6%–2.3% if the maximum number of SS/PBCH blocks is transmitted [73,74].

4.1.2.1 Demodulation Reference Signals

The main application of DM-RS in NR is to estimate the channel coefficients for coherent detection of the physical channels. In the downlink, the DM-RS is used for channel estimation and is subject to the same precoding as PDSCH; thus the (transmit-side) precoding is transparent to the receiver and is viewed as part of the overall channel. There is a trade-off between the channel estimation accuracy and DM-RS density/overhead. If the channel exhibits severe frequency-selectivity (i.e., narrower channel coherence bandwidth), the DM-RS density in the frequency-domain should be increased. Similarly, if the channel varies faster in time-domain (i.e., shorter channel coherence time), denser DM-RS allocation across time is required. After determining frequency/time-domain DM-RS densities, the DM-RS locations in the time-frequency resource grid should be considered. Assuming stationary channel conditions, uniform DM-RS allocation in both frequency and time-domain is preferred for minimizing interpolation error and reducing implementation complexity. Since no user data is transmitted by DM-RS per se, allocating DM-RS with a proper density is required to maximize the throughput.

In NR, a front-loaded DM-RS structure is used as a baseline to achieve low-latency decoding (see Fig. 4.2). In the time-frequency resource grid, the front-loaded DM-RS can be located just after the control region, followed by data region. As soon as channel is

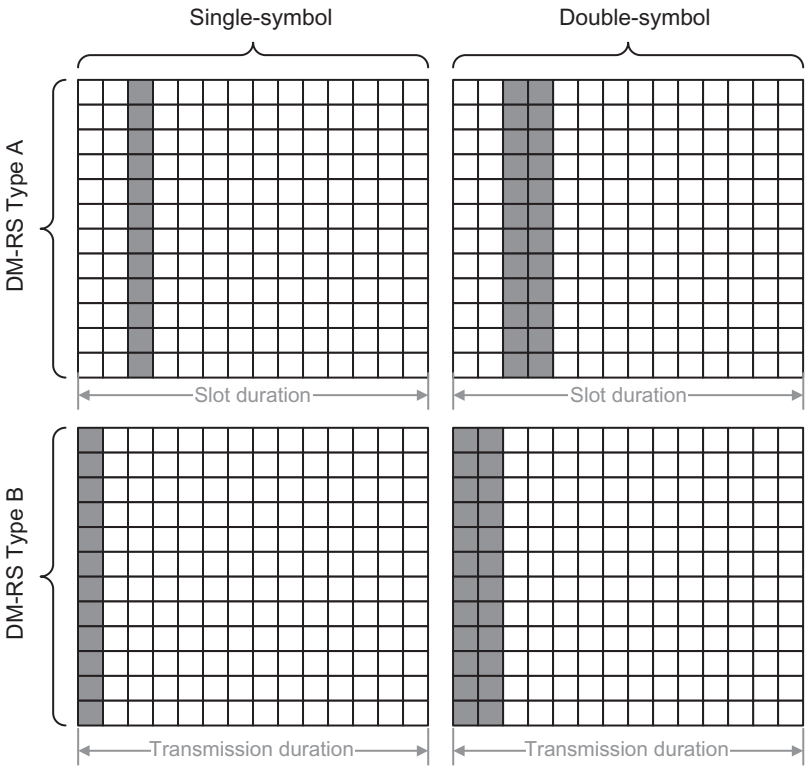


Figure 4.2
Comparison of DM-RS Type A and Type B mappings [6].

estimated based on the front-loaded DM-RS, the receiver can coherently demodulate data in the data region. The front-loaded DM-RS structure is particularly advantageous in decoding-latency reduction for low-mobility scenarios where channel coherence time is longer than the duration of the front-loaded DM-RS. However, allocating only the front-loaded DM-RS can degrade the link performance at higher UE speeds (i.e., channel coherence time becomes shorter). Although the channel information in the data region can be obtained by interpolation, the channel information accuracy diminishes with higher mobility. Therefore, we consider the front-loaded DM-RS patterns with $2\times$ and $4\times$ time-domain densities as shown in Fig. 4.3 [70]. To support high-speed scenarios, it is possible to configure up to three additional DM-RS occasions in a slot. The channel estimation in the receiver side can use these additional reference signals for more accurate channel estimation, for example, to perform interpolation between the DM-RS occasions within a slot. However, unlike LTE, it is not possible to interpolate channel estimations between slots, or in general different transmission occasions, since different slots may be transmitted to different devices and/or in different beam directions [14].

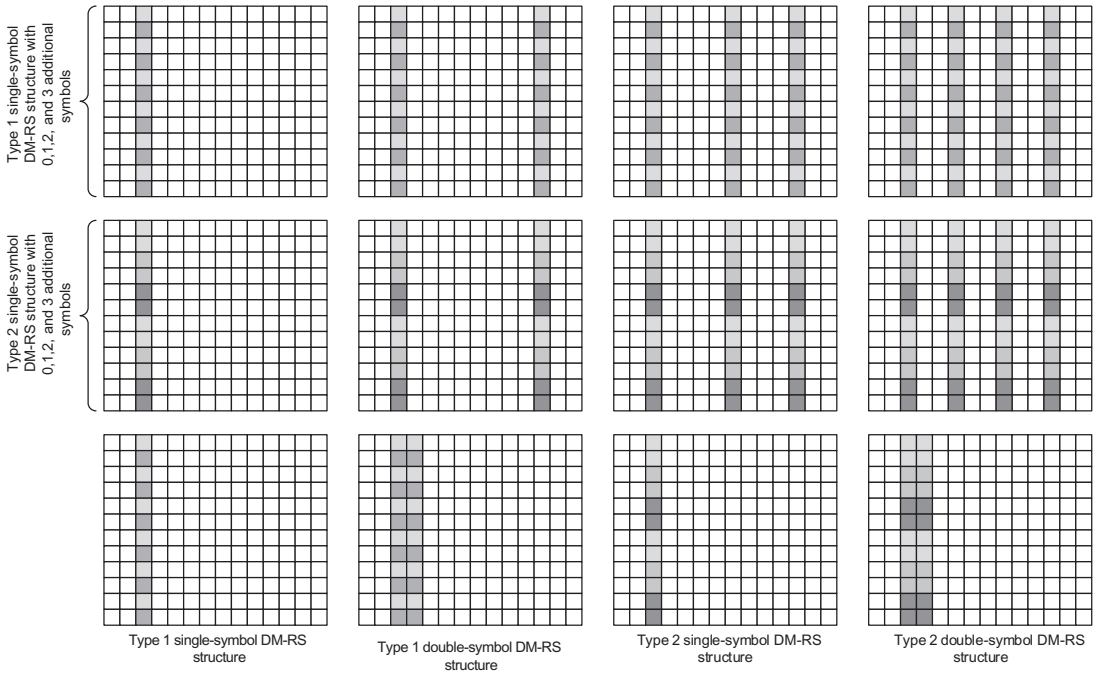


Figure 4.3
Various NR PDSCH DM-RS Type A time-frequency patterns [6].

In LoS-dominant channel conditions, the delay spread is expected to be shorter (or equivalently the channel coherence bandwidth becomes larger); thus one can consider reducing frequency-domain density of the DM-RS without significant degradation of channel estimation precision. By doing so, the overhead due to the DM-RS can be reduced. One example of such low-density DM-RS patterns in frequency-domain is shown in Fig. 4.3. For MIMO transmission up to two frequency-domain orthogonal DM-RS ports are supported. The DM-RS is UE-specific, can be beamformed, is confined in the UE scheduled resources, and is transmitted only when necessary, both in the downlink and uplink and is used to estimate the communication channel prior to coherent demodulation. To support multi-layer MIMO transmission, multiple orthogonal DM-RS ports can be scheduled, one for each layer. Orthogonality is achieved by means of FDM (comb structure), TDM, and/or CDM (with cyclic shift of the base sequence or orthogonal cover codes) methods. The basic DM-RS pattern is front loaded, as the DM-RS design considers the early decoding requirement to support low-latency applications. For low-speed scenarios, DM-RS uses low density in the time domain. However, for high-speed scenarios, the time density of DM-RS is increased to track fast changes in the

radio channel. The NR defines two time-domain DM-RS structures which differ in the location of the first DM-RS symbol [14]:

- *Mapping Type A*, where the first DM-RS is located in the second and the third symbol of the slot and the DM-RS is mapped relative to the start of the slot boundary, regardless of where in the slot the actual data transmission occurs. This mapping type is primarily intended for the case where the data occupy (most of) a slot. The reason for the use of the second or the third symbol in the downlink slot is to locate the first DM-RS occasion after a CORESET that is positioned at the beginning of a slot (see Fig. 4.2).
- *Mapping Type B*, where the first DM-RS is positioned in the first symbol of the data allocation, that is, the DM-RS location is not given relative to the slot boundary, rather relative to where the data are located. This mapping is intended for transmissions over a small fraction of the slot to support very low latency and other transmissions that cannot wait until a slot boundary starts regardless of the transmission duration. The mapping type for PDSCH transmission can be dynamically signaled as part of the downlink control information (DCI), while for the physical uplink shared channel (PUSCH) the mapping type is semi-statically configured (see Fig. 4.2).

The different time-domain locations for (PDSCH) DM-RS mapping types are illustrated in Figs. 4.2 and 4.3, including both single-symbol and double-symbol DM-RS patterns. The purpose of the double-symbol DM-RS is primarily to provide a larger number of antenna ports than what is possible with a single-symbol structure as discussed later. Note that the time-domain location of the DM-RS depends on the scheduled data duration. Multiple orthogonal reference signals can be generated in each DM-RS occasion.

Different DM-RS patterns can be configured which are separated in time, frequency, and code domains. The DM-RS has two types: that is, Types 1 and 2, which are distinguished in frequency-domain mapping and the maximum number of orthogonal reference signals. Type 1 can provide up to four orthogonal signals using a single-symbol DM-RS and up to eight orthogonal reference signals using a double-symbol DM-RS, whereas Type 2 can provide 6 and 12 patterns depending on the number of symbols. The DM-RS Type 1 or 2 should not be confused with the mapping Type A or B, since different mapping types can be combined with different reference signal types. Reference signals should preferably have small power variations in the frequency domain to allow a similar channel-estimation quality for all frequencies spanned by the reference signal. Note that this is equivalent to a highly localized time-domain autocorrelation of the transmitted reference signal.

The PDSCH DM-RS sequence $r_{DM\text{-}RS}(n)$ is defined as $r_{DM\text{-}RS}(n) = \{[1 - 2c(2n)] + j[1 - 2c(2n+1)]\}/\sqrt{2}$, where $c(i)$ is a length-31 Gold sequence[1] generated by the pseudo-random sequence generator defined in [6] and initialized with $c_{init} = [2^{17}(N_{slot}^{symbol} n_{slot} + l + 1)(2N_{ID}^{n_{SCID}} + 1) + 2N_{ID}^{n_{SCID}} + n_{SCID}] \bmod 2^{31}$. In the latter expression, l denotes the OFDM symbol number within the slot, n_{slot} is the slot number within a frame, and $n_{SCID} = \{0, 1\}$ is given by the DM-RS sequence initialization field in the DCI associated with the PDSCH transmission, if DCI format 1_1 is used and $N_{ID}^{n_{SCID}} = \{0, 1, \ldots, 65535\}$ is the scrambling identifier when signaled by higher layer parameters *scramblingID0* and *scramblingID1*; otherwise $n_{SCID} = 0$ and $N_{ID}^{n_{SCID}} = N_{ID}^{cell}$. The PDSCH is scheduled by PDCCH using DCI format 1_1 (or DCI format 1_0) with CRC scrambled by C-RNTI, MCS-C-RNTI, or CS-RNTI [6].

The PDSCH DM-RS can be mapped to physical resources in two ways referred to as configuration Type 1 or 2, which is determined by RRC parameter *dmrs-Type*. Prior to resource mapping, the sequence $r_{DM\text{-}RS}(m)$ is multiplied by scaling factor β_{PDSCH}^{DMRS} to adjust the transmission power and is mapped to resource element (RE) (k, l) as $\alpha(k, l) = \beta_{PDSCH}^{DMRS} w_f(k') w_t(l') r_{DM\text{-}RS}(2n + k')$ where $w_f(k')$ and $w_t(l')$ are orthogonal cover codes or spreading functions across frequency and time that are given in [6]. In the latter expression, $n \in \mathbb{N}$, $l = \bar{l} + l'$, and $k' = 0, 1$ and $k = 4n + 2k' + \Delta$ for configuration Type 1 and $k = 6n + k' + \Delta$ for configuration Type 2.

In DM-RS Type 1, the underlying pseudo-random sequence is mapped to every other subcarrier in the frequency domain over the OFDM symbol used for reference signal transmission (see Fig. 4.3). Antenna ports 1000 and 1001 use even-numbered subcarriers in the frequency domain and are separated from each other by multiplying the underlying pseudo-random sequence with different length-2 orthogonal sequences in the frequency domain, resulting in transmission of two orthogonal reference signals for the two antenna ports. If the radio channel can be considered flat across four consecutive subcarriers, the two reference signals will maintain orthogonality at the receiver. Antenna ports 1000 and 1001 are said to belong to CDM group 0, since they use the same subcarriers but are separated in the code-domain using different orthogonal sequences. Reference signals for antenna ports 1002 and 1003 belong to CDM group 1 and are generated in the same way using odd-numbered

[1] A Gold sequence is a type of binary sequence that is often used in telecommunication and satellite navigation. Gold sequences have bounded small cross-correlations within a set. A set of Gold sequences consists of $2^n - 1$ sequences each with a period of $2^n - 1$. A set of Gold sequences can be generated by taking two maximum-length sequences of the same length $2^n - 1$ such that their absolute cross-correlation is less than or equal to $2^{(n+2)/2}$, where n is the size of the linear feedback shift register used to generate the maximum length sequence. The set of $2^n - 1$ (logical) exclusive OR of the two sequences in their various phases is a set of Gold codes. The highest absolute cross-correlation in this set of codes is $2^{(n+2)/2} + 1$ for even n, and $2^{(n+1)/2} + 1$ for odd n. The exclusive OR of two Gold codes from the same set is another Gold code with arbitrary phase.

subcarriers and are separated in the code domain within the CDM group and in the frequency domain between CDM groups. If more than four orthogonal antenna ports are needed, two consecutive OFDM symbols are used instead. The aforementioned structure is used over each of the OFDM symbols and a length-2 orthogonal sequence is used to extend the code-domain separation over time, resulting in up to eight orthogonal sequences.

The DM-RS Type 2 has a similar structure to Type 1, except some differences with to the number of antenna ports that are supported. Each CDM group for Type 2 consists of two neighboring subcarriers over which a length-2 orthogonal sequence is used to separate the two antenna ports sharing the same set of subcarriers. Four subcarriers are used in each resource block and in each CDM group. Since there are 12 subcarriers in a resource block, up to three CDM groups with two orthogonal reference signals can be created using one resource block over one OFDM symbol. If a second OFDM symbol is used along with a length-2 sequence in time-domain, up to 12 orthogonal reference signals can be generated [14].

The location of front-loaded DM-RS symbols, which can be either one or two symbols, is dependent on whether a slot based (DM-RS mapping Type A) or non-slot-based (DM-RS mapping Type B) scheduling is used. In the former type, fixed OFDM symbols regardless of the PDSCH assignment are used to map DM-RS (configurable via parameter $l_0 = \{2, 3\}$), whereas in latter type which corresponds to mini-slots, the first OFDM symbol assigned for PDSCH is used to map DM-RS. The reference point for l and the position l_0 of the first DM-RS symbol depends on the mapping type. Additional DM-RS symbols can be configured (e.g., for high-speed scenarios) as well as for broadcast/multicast PDSCH. In the preceding equations, the reference point for frequency index k depends on PDSCH payload. The reference point for frequency index k is subcarrier 0 of the lowest numbered resource block in CORESET 0, if the corresponding PDCCH is associated with CORESET 0 and Type0-PDCCH common search space and identified by system information (SI)-RNTI; otherwise, it is the subcarrier 0 in common resource block 0. Furthermore, the reference point for time index l and the reference position l_0 of the first DM-RS symbol depends on the mapping type, which for PDSCH mapping Type A, l is defined relative to the start of the slot, that is, $l_0 = 3$ if the RRC parameter *dmrs-TypeA-Position* equals 3; otherwise, $l_0 = 2$ and for PDSCH mapping Type B, l is defined relative to the start of the scheduled PDSCH resources, that is, $l_0 = 0$. The position of the DM-RS symbols is further dependent on parameter $\bar{l}$ where for PDSCH mapping Type A, the duration is between the first OFDM symbol of the slot, and the last OFDM symbol of the scheduled PDSCH resources in the slot; and for PDSCH mapping Type B, the duration is the number of OFDM symbols of the scheduled PDSCH resources given by the parameters specified in [6].

For PDSCH mapping Type B, if the PDSCH duration is 2, 4, or 7 OFDM symbols (i.e., mini-slot scheduling), and if the PDSCH allocation collides with resources reserved for a

CORESET, $\bar{l}$ is incremented such that the first DM-RS symbol is located immediately following the CORESET. If PDSCH duration is 2, 4, or 7 symbols, the UE would not expect to receive a DM-RS symbol beyond the second, third, and fourth symbol, respectively. If one additional single-symbol DM-RS is configured, the UE expects the additional DM-RS to be transmitted on the fifth or sixth symbol when the front-loaded DM-RS symbol is in the first or second symbol, respectively; otherwise, the UE should expect that the additional DM-RS is not transmitted. If PDSCH duration is two or four OFDM symbols, only a single-symbol DM-RS is supported. Furthermore, single-symbol or double-symbol DM-RS is used, if RRC parameter *maxLength* is equal to 1 or 2, respectively [8,9].

In the absence of CSI-RS configuration, the UE can assume PDSCH DM-RS and SS/PBCH block antenna ports are quasi-co-located with respect to Doppler shift, Doppler spread, average delay, delay spread, and spatial RX[2] parameters. The UE may assume that the PDSCH DM-RS within the same CDM group are quasi-co-located with respect to Doppler shift, Doppler spread, average delay, delay spread, and spatial RX parameters. Note that the spatial RX parameters are meant to describe angular/spatial channel properties at the UE to help the UE select and use one of the beams. The UE can use SS/PBCH block to obtain frequency offset, timing offset, Doppler spread, delay spread, and receive beam to process DM-RS. In other words, one can consider the spatial RX parameters as beam indication for the UE, where UE may use the acquired channel parameters from SS/PBCH to receive PDSCH.

4.1.2.2 Phase Tracking Reference Signals

The PT-RS was introduced in NR to enable compensation of oscillator phase noise in above-6 GHz frequency bands. Phase noise typically increases as a function of carrier frequency. Therefore, PT-RS can be utilized at high carrier frequencies (e.g., mmWave bands) to mitigate the phase noise effect. In the case of OFDM signals, the effect of phase noise is identical phase rotation of all the subcarriers, known as CPE. In NR, the PT-RS is designed

[2] To explain the spatial RX concept, let's consider receive antenna diversity where the transmitted signal is received by N antennas which are assumed to have sufficient spatial separation resulting in independent multipath channels between the transmitter and each receiver antenna. Denoting the channel vector, including the amplitude and phase coefficients representing the aggregated effects of multipath propagation, by $\mathbf{h} = (h_1, h_2, \ldots, h_N) \in \mathbb{C}^{N \times 1}$, the received baseband equivalent signal vector $\mathbf{r} \in \mathbb{C}^{N \times 1}$ can be expressed by $\mathbf{r} = \mathbf{hx} + \mathbf{n}$. Consequently, by combining the signals from the separate antenna branches with a proper weighting, we obtain the combiner output being equal to $y_{RX} = \mathbf{W}^H\mathbf{r} = \mathbf{W}^H\mathbf{hx} + \mathbf{W}^H\mathbf{n}$, where $\mathbf{W} \in \mathbb{C}^{N \times 1}$ denotes the weighting vector of the combiner. It is important to note that now the steering vector is replaced with a more generic channel response, including arbitrary amplitude and phase response for each antenna branch. Consequently, the combiner weights do not match with any particular physical direction anymore, rather they need to be adjusted according to the generic channel response and this form of multi-antenna-based signal combining is referred to as spatial RX processing. The actual weight selection and optimization, in turn, can be implemented by several methods, which differ in complexity and performance. The simplest method is selection combining where only the signal with the highest instantaneous SNR is used for detection.

so that it has low density in the frequency domain and high density in the time domain, because the phase rotation caused by CPE is identical for all subcarriers within an OFDM symbol; however, it has minimal correlation across OFDM symbols. The PT-RS is UE-specific, confined in a scheduled resource, and can be beamformed. The number of PT-RS ports can be lower than the total number of ports, and orthogonality between PT-RS ports is achieved by means of frequency-division multiplexing. The PT-RS is configurable depending on the quality of the oscillators, carrier frequency, OFDM subcarrier spacing, and modulation and coding schemes used for transmission.

The PT-RS introduced in NR is used for time and frequency tracking as well as estimation of delay spread, and Doppler spread at the UE side. They are transmitted in a confined bandwidth for a configurable time duration controlled by RRC parameters. The time-frequency structure of PT-RS depends on the waveform. For OFDM, the first reference symbol (prior to applying any orthogonal sequence) in a PDSCH/PUSCH allocation is repeated every $L_{PT\text{-}RS} \in \{1, 2, 4\}$ symbol, starting with the first OFDM symbol in the allocation. The repetition counter is reset at each DM-RS position since there is no need for PT-RS insertion immediately following a DM-RS occasion. In the frequency domain, PT-RS are transmitted in every second or fourth resource block, resulting in a sparse frequency-domain structure. The density in the frequency domain is dependent on the scheduled bandwidth in a sense that the higher the bandwidth, the lower the PT-RS density. For the smallest bandwidths, no PT-RS is transmitted. To reduce the risk of collision between PT-RS associated with different devices scheduled on overlapping frequency-domain resources, the subcarrier number and the resource blocks used for PT-RS transmission are determined by the C-RNTI of the device. The antenna port used for PT-RS transmission is given by the lowest numbered antenna port in the DM-RS antenna port group. An example time-frequency PT-RS structure is shown in Fig. 4.4.

The PT-RS for subcarrier k is given by $r_{PT\text{-}RS}(k) = r_{DM\text{-}RS}(2m + k')$ where $r_{DM\text{-}RS}(2m + k')$ is the DM-RS at time-domain position l_0 and subcarrier k. The PT-RS is present only in the resource blocks used for the PDSCH, and only if there is an explicit indication of their presence. In that case, the PT-RS is scaled by a factor of $\beta_{PT\text{-}RS}(i)$ to adjust the transmission power. The PT-RS is mapped to resource elements $a(k, l) = \beta_{PT\text{-}RS}(i) r_{PT\text{-}RS}(k)$, if l is located within the OFDM symbols allocated for the PDSCH transmission and the designated resource element is not used for DM-RS, CSI-RS, SS/PBCH block, PDCCH, or is declared as not available. The time indices l at which PT-RS are allocated are defined relative to the start of the PDSCH allocation l_0 and are given by $l = l_0 + iL_{PT\text{-}RS}$ where i is incremented as long as the PT-RS occasion falls inside the PDSCH allocation and the aforementioned conditions are met. For PT-RS resource mapping, the resource blocks allocated for PDSCH transmission are numbered from 0 to $N_{RB} - 1$ from the lowest scheduled resource block to the highest. The corresponding subcarriers in this set of resource blocks are numbered in increasing order starting from the lowest frequency to $N_{sc}^{RB} N_{RB} - 1$. The subcarrier indices that the PT-RS are

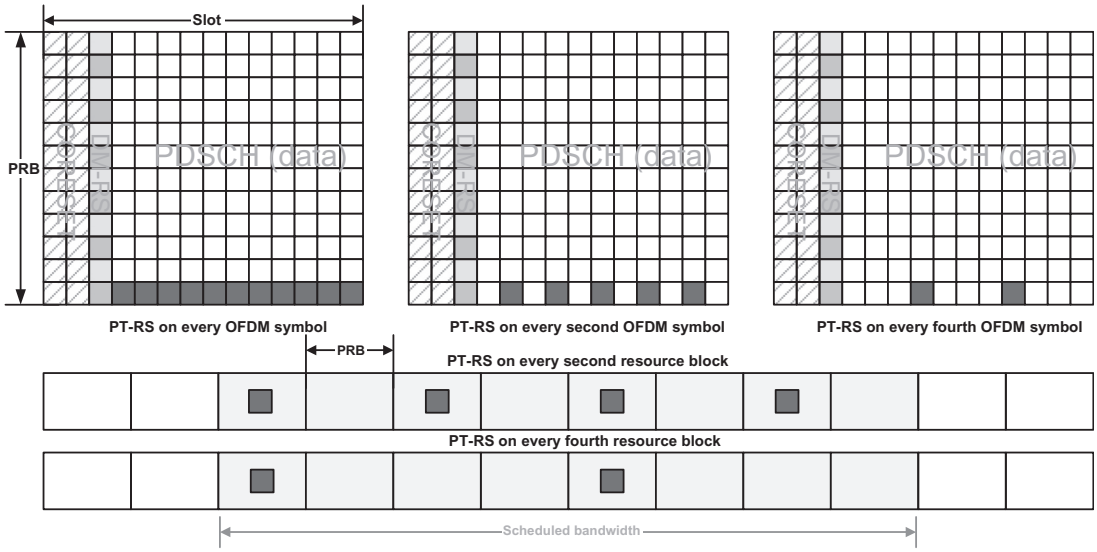


Figure 4.4
Illustration of PT-RS structures in time and frequency domain [6].

mapped to are given by $k = k_{ref}^{RE} + (iK_{PTRS} + k_{ref}^{RB})N_{sc}^{RB}; i \in \mathbb{N}$ where $k_{ref}^{RB} = n_{RNTI} \bmod K_{PT\text{-}RS}$ if $N_{RB} \bmod K_{PT\text{-}RS} = 0$; otherwise, $k_{ref}^{RB} = n_{RNTI} \bmod (N_{RB} \bmod K_{PT\text{-}RS})$ [6]. In the latter equation, n_{RNTI} denotes the RNTI associated with the DCI scheduling of the transmission; $K_{PT-RS} \in \{2, 4\}$, and k_{ref}^{RE} is the DM-RS port associated with the PT-RS port. The density of PT-RS in time and frequency domain is configurable to address different scenarios (e.g., different carrier frequency, modulation and coding scheme, and hardware quality). The PT-RS patterns in time and frequency domains are illustrated in Fig. 4.4.

4.1.2.3 Channel State Information Reference Signals

The CSI-RS in NR is used for downlink CSI estimation. It further supports RSRP measurements for mobility and beam management (including analog beamforming), time/frequency tracking for demodulation, and uplink reciprocity-based precoding. The CSI-RS is UE-specific; nevertheless, multiple users can share the same CSI-RS resource. The NR defines zero-power and non-zero-power CSI-RS. When a zero-power CSI-RS is configured, the resource elements (designated to CSI-RS) are not used for PDSCH transmission. In this case, the zero-power CSI-RS is used to mask certain resource elements, making them unavailable for PDSCH mapping. This masking not only supports transmission of UE-specific CSI-RS, but also the design allows introduction of new features while maintaining backward compatibility. The NR supports flexible CSI-RS configurations. A CSI resource can be configured with up to 32 antenna ports with configurable density. In the time domain, a CSI-RS resource may start at any OFDM symbol of a slot and span 1, 2, or 4

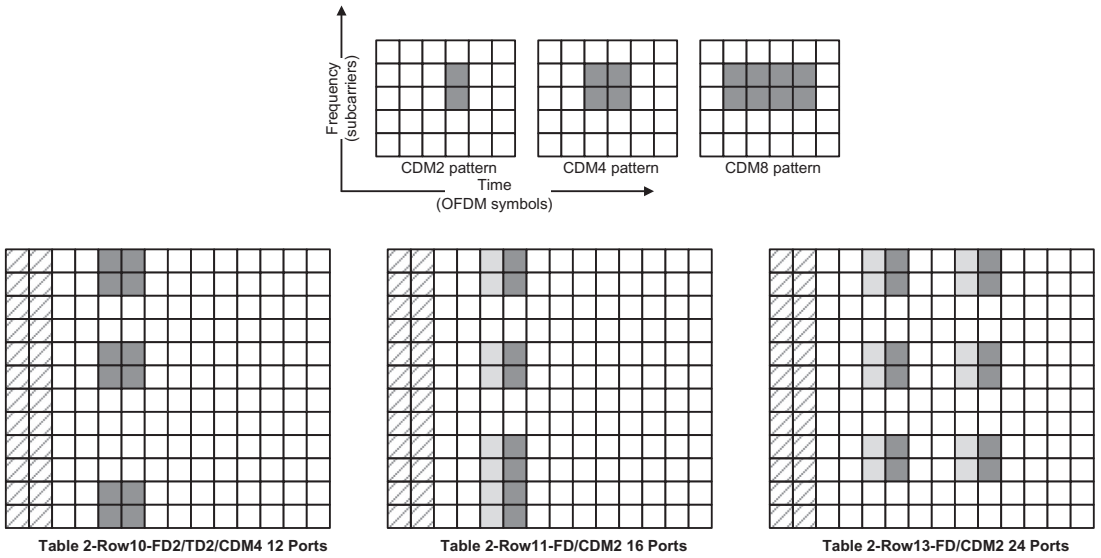


Figure 4.5
Example locations of CSI-RS in time and frequency [6].

OFDM symbols depending on the number of configured antenna ports. The CSI-RS can be periodic, semi-persistent, or aperiodic (DCI triggered). When used for time frequency tracking, the CSI-RS can be periodic or aperiodic. In this case, a single port is configured, and the signal is transmitted in the form of bursts over two or four symbols that are spread over one or two slots [71].

A configured CSI-RS resource may correspond to up to 32 different antenna ports. In NR, a CSI-RS is always configured on a per-device basis. It must be noted that the UE-specific configuration of CSI-RS does not necessarily mean that a transmitted CSI-RS can only be used by a single device, rather the same set of CSI-RS resources can be separately configured for multiple devices, which means that a single CSI-RS can be shared among multiple devices. As illustrated in Fig. 4.5, a single-port CSI-RS occupies a single resource element within a resource block in the frequency domain and one slot in the time domain. While the CSI-RS can be configured to occur anywhere within the resource block, in practice there are some restrictions on CSI-RS resource assignment to avoid collisions with other downlink physical channels and signals. The transmission of a configured CSI-RS is expected not to collide with a CORESET configured for the device; the DM-RS associated with PDSCH transmissions scheduled for the device; and the SS blocks transmissions.

A multiport CSI-RS can be viewed as multiple orthogonal per antenna-port CSI-RS sharing the set of resource elements assigned for transmission of the configured multiport CSI-RS. In general, the resource sharing is achieved through a combination of code-domain

(i.e.,different per antenna-port CSI-RS are transmitted on the same set of resource elements with separation achieved by spreading the CSI-RS with different orthogonal codes), frequency-domain (i.e., different per antenna-port CSI-RS are transmitted on different subcarriers over an OFDM symbol), and time-domain (i.e., different per antenna-port CSI-RS are transmitted on different OFDM symbols within a slot) multiplexing schemes. As illustrated in Fig. 4.5, code-division multiplexing between different (per antenna-port) CSI-RS can be performed across frequency by spreading over two adjacent subcarriers (CDM2) to support two antenna ports; across frequency and time by spreading over two adjacent subcarriers and two adjacent OFDM symbols (CDM4) to enable four antenna ports; across frequency and time by spreading over two adjacent subcarriers and four adjacent OFDM symbols (CDM8) to support up to eight antenna port transmission. Combination of code/frequency and time-division multiplexing can be used to configure different multiport CSI-RS structures where, in general, an N-port CSI-RS occupies N resource elements within a resource block or a slot. When CSI-RS supports more than two antenna ports, there are multiple CSI-RS patterns/structures based on different combinations of CDM, TDM, and FDM that can be utilized as shown in Table 4.2.

The non-zero-power CSI-RS is mathematically represented by sequence $r_{CSI\text{-}RS}(m)$ which is defined as $r_{CSI\text{-}RS}(m) = \{[1 - 2c(2m)] + j[1 - 2c(2m+1)]\}/\sqrt{2}$, where $c(i)$ is a length-31 Gold sequence generated by the pseudo-random sequence generator defined in [6] and initialized with $c_{init} = [2^{10}(N_{slot}^{symbol} n_{slot} + l + 1)(2n_{ID} + 1) + n_{ID}] \bmod 2^{31}$ at the start of each OFDM symbol. In the latter equation, n_{slot} is the slot number within a radio frame, l is the OFDM symbol number within a slot, and n_{ID} is set by the RRC parameter *scramblingID* or *sequenceGenerationConfig* [6,13].

If CSI-RS is configured, the CSI-RS sequence $r_{CSI\text{-}RS}(m)$ is mapped to resources elements (k, l) such that $a(k, l) = \beta_{CSI\text{-}RS} w_f(k') w_t(l') r_{CSI\text{-}RS}(m')$ where $m' = \lfloor n\alpha \rfloor + k' + \left\lfloor k\rho / N_{sc}^{RB} \right\rfloor \forall n \in \mathbb{N}, k = nN_{sc}^{RB} + k'$ and $l = \bar{l} + l'$, $\alpha = \rho$ or 2ρ when $N_p = 1$ or $N_p > 1$, respectively, provided that the resource element (k, l) is within the resource blocks designated to the CSI-RS resource for which the UE is configured. The reference point for $k = 0$ is subcarrier 0 in common resource block 0. The value of density ρ and the number of antenna ports are given by RRC parameters [6,13]. The scaling factor $\beta_{CSI\text{-}RS}$ has a non-zero value for a non-zero-power CSI-RS to ensure that the power offset specified by the RRC parameter *powerControlOffsetSS* is satisfied. Other parameters $k', l', w_f(k')$, and $w_t(l')$ are given in Table 4.2 where each pair $(\bar{k}, \bar{l})$ corresponds to a CDM group of size 1 (no CDM) or size 2, 4, or 8. The indices k' and l' are used to index the resource elements within a CDM group [6,8]. The time-domain locations $l_0 = \{0, 1, \ldots, 13\}$ and $l_1 = \{2, 3, \ldots, 12\}$ are defined relative to the start of a slot with the starting positions of a CSI-RS in a slot configured by the RRC parameters provided in *CSI-RS-ResourceMapping* information element. The frequency-domain location of CSI-RS is determined by a bitmap signaled via the RRC parameter provided in

Table 4.2: Various CSI-RS patterns (locations) within a slot [6].

Row	Number of Ports N_p	CSI-RS Density ρ	CDM Type	$(\bar{k}, \bar{l})$	CDM Group Index j	k'	l'
1	1	3	No CDM	$(k_0, l_0), (k_0+4, l_0), (k_0+8, l_0)$	0,0,0	0	0
2	1	1,0.5	No CDM	(k_0, l_0)	0	0	0
3	2	1,0.5	FD-CDM2	(k_0, l_0)	0	0,1	0
4	4	1	FD-CDM2	$(k_0, l_0), (k_0+2, l_0)$	0,1	0,1	0
5	4	1	FD-CDM2	$(k_0, l_0), (k_0, l_0+1)$	0,1	0,1	0
6	8	1	FD-CDM2	$(k_0, l_0), (k_1, l_0), (k_2, l_0), (k_3, l_0)$	0,1,2,3	0,1	0
7	8	1	FD-CDM2	$(k_0, l_0), (k_1, l_0), (k_0, l_0+1), (k_1, l_0+1)$	0,1,2,3	0,1	0
8	8	1	CDM4 (FD2,TD2)	$(k_0, l_0), (k_1, l_0)$	0,1	0,1	0,1
9	12	1	FD-CDM2	$(k_0, l_0), (k_1, l_0), (k_2, l_0), (k_3, l_0), (k_4, l_0), (k_5, l_0)$	0,1,2,3,4,5	0,1	0
10	12	1	CDM4 (FD2,TD2)	$(k_0, l_0), (k_1, l_0), (k_2, l_0)$	0,1,2	0,1	0,1
11	16	1,0.5	FD-CDM2	$(k_0, l_0), (k_1, l_0), (k_2, l_0), (k_3, l_0)$ $(k_0, l_0+1), (k_1, l_0+1), (k_2, l_0+1), (k_3, l_0+1)$	0,1,2,3,4,5,6,7	0,1	0
12	16	1,0.5	CDM4 (FD2,TD2)	$(k_0, l_0), (k_1, l_0), (k_2, l_0), (k_3, l_0)$	0,1,2,3	0,1	0,1
13	24	1,0.5	FD-CDM2	$(k_0, l_0), (k_1, l_0), (k_2, l_0), (k_0, l_0+1), (k_1, l_0+1), (k_2, l_0+1)$ $(k_0, l_1), (k_1, l_1), (k_2, l_1), (k_0, l_1+1), (k_1, l_1+1), (k_2, l_1+1)$	0,1,2,3,4,5,6,7,8,9,10,11	0,1	0
14	24	1,0.5	CDM4 (FD2,TD2)	$(k_0, l_0), (k_1, l_0), (k_2, l_0), (k_0, l_1), (k_1, l_1), (k_2, l_1)$	0,1,2,3,4,5	0,1	0,1
15	24	1,0.5	CDM8 (FD2,TD4)	$(k_0, l_0), (k_1, l_0), (k_2, l_0)$	0,1,2	0,1	0,1,2,3
16	32	1,0.5	FD-CDM2	$(k_0, l_0), (k_1, l_0), (k_2, l_0), (k_3, l_0),$ $(k_0, l_0+1), (k_1, l_0+1), (k_2, l_0+1), (k_3, l_0+1)$ $(k_0, l_1), (k_1, l_1), (k_2, l_1), (k_3, l_1),$ $(k_0, l_1+1), (k_1, l_1+1), (k_2, l_1+1), (k_3, l_1+1)$	0,1,2,3, 4,5,6,7, 8,9,10,11, 12,13,14,15	0,1	0
17	32	1,0.5	CDM4 (FD2,TD2)	$(k_0, l_0), (k_1, l_0), (k_2, l_0), (k_3, l_0),$ $(k_0, l_1), (k_1, l_1), (k_2, l_1), (k_3, l_1)$	0,1,2,3,4,5,6,7	0,1	0,1
18	32	1,0.5	CDM8 (FD2,TD4)	$(k_0, l_0), (k_1, l_0), (k_2, l_0), (k_3, l_0)$	0,1,2,3	0,1	0,1,2,3

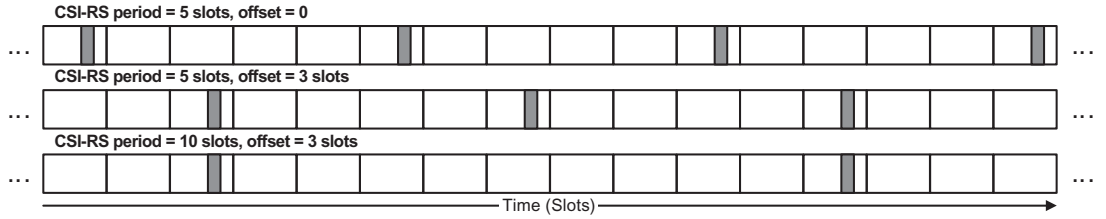


Figure 4.6
Example CSI-RS periodicity and offset [6].

CSI-RS-ResourceMapping information element [6,13]. The starting position and number of the resource blocks in which the CSI-RS is transmitted are provided via RRC signaling.

The CSI-RS is transmitted on antenna ports $p = 3000 + s + jL_{CDM}$ where $j = 0, 1, \ldots, N_p/L_{CDM} - 1$ and $s = 0, 1, \ldots, L_{CDM} - 1$. In the latter expression, $L_{CDM} \in \{1, 2, 4, 8\}$ is the CDM group size and N_p is the number of CSI-RS antenna ports. The CDM groups are numbered in order of increasing frequency-domain allocation first and then increasing time domain-allocation. For a CSI-RS resource configured as periodic or semi-persistent by the RRC parameter *resourceType*, the CSI-RS is transmitted in slot numbers satisfying $(N_{frame}^{slot} n_{slot} + n_{slot} - T_{offset}) \bmod T_{CSI\text{-}RS} = 0$ where the CSI-RS periodicity $T_{CSI\text{-}RS}$ (in number of slots) and slot offset T_{offset} are signaled by the RRC parameter *CSI-ResourcePeriodicityAndOffset* or *slotconfig* (see Fig. 4.6). The CSI-RS is transmitted in a slot only if, all OFDM symbols of that slot corresponding to the configured CSI-RS resource are designated for downlink transmission. The antenna ports within a CSI-RS resource are quasi-co-located with quasi-co-location (QCL) Type A, Type D (when applicable). In summary, the NR supports periodic, aperiodic, semi-persistent CSI-RS transmission. The NR CSI-RS patterns can be mapped to 1, 2, or 4 OFDM symbols and support CDM2, CDM4, CDM8 spreading functions, for example, CDM8 means that there are eight spreading functions $w_f(k')$ and $w_t(l')$. For CSI acquisition, the NR supports CSI-RS density $\rho = 0.5$ and 1 RE/RB/port and a PRB-level comb-type transmission. The number of antenna ports can be independently configured for periodic, aperiodic, semi-persistent CSI reporting. A CSI-RS resource configuration up to 32 ports is supported in NR. The UE-specific CSI-RS may be configured to support wideband CSI-RS and partial band CSI-RS. In order to reduce beam management overhead and latency, the NR supports subtime units of less than one OFDM symbol in a reference numerology [73]. Each CSI-RS resource is configured by the RRC parameter *NZP-CSI-RS-Resource*. The time-domain locations of the two periodic CSI-RS resources in a slot or four periodic CSI-RS resources in two consecutive slots are given by $l \in \{4, 8\}, l \in \{5, 9\}$, or $l \in \{6, 10\}$ for FR1 and FR2; or $l \in \{0, 4\}, l \in \{1, 5\}, l \in \{2, 6\}, l \in \{3, 7\}, l \in \{7, 11\}, l \in \{8, 12\}$ or $l \in \{9, 13\}$ for FR2. A single-port

CSI-RS resource with density $\rho = 3$ (see Table 4.2) and RRC parameter *density* is configured by *CSI-RS-ResourceMapping*. The bandwidth of the CSI-RS resource, given by RRC parameter *freqBand* configured by *CSI-RS-ResourceMapping*, is determined as $\min(52, N_{RB}^{BWP(i)})$ resource blocks or is equal to $N_{RB}^{BWP(i)}$ resource blocks. The UE is not expected to be configured with the periodicity of $2^{\mu} \times 10$ slots, if the bandwidth of CSI-RS resource is larger than 52 resource blocks. The periodicity and slot offset, given by RRC parameter *periodicityAndOffset* configured by *NZP-CSI-RS-Resource* parameter, is one of $2^{\mu}X_p$ slots where $X_p = 10, 20, 40,$ or 80 [9]. It should be noted that the property of periodic, semi-persistent, or aperiodic is not a property of the CSI-RS per se, rather the property of a CSI-RS resource set. As a result, activation/deactivation and triggering of semi-persistent and aperiodic CSI-RS must be done for a set of CSI-RS within a resource set. In the case of periodic CSI-RS transmission, the UE can assume that a configured CSI-RS transmission occurs every *N*th slot, where *N* ranges from 4 to 640. In addition to the periodicity, the device is also configured with a specific slot offset for the CSI-RS transmission. In the case of semi-persistent CSI-RS transmission, certain CSI-RS periodicity and slot offset are configured similar to periodic CSI-RS transmission. However, the CSI-RS transmission can be activated or deactivated via MAC control elements. Once the CSI-RS transmission has been activated, the device can assume that the CSI-RS transmission will continue according to the configured periodicity until it is deactivated. Similarly, once the CSI-RS transmission has been deactivated, the device can assume that there will be no CSI-RS transmission according to the configuration until it is reactivated. In the case of aperiodic CSI-RS, no periodicity is configured, rather the UE is triggered via signaling in the DCI [9].

The CSI-RS may be further used for RSRP measurements[3] and mobility management since NR does not include the cell-specific reference signals that were used in LTE for mobility management. The set of CSI-RS corresponding to a set of beams on which measurements are conducted should be included in the non-zero power (NZP)-CSI-RS resource set associated with the report configuration. Such a resource set may either include a set of configured CSI-RS or a set of SS blocks. Measurements for beam management can be carried out on either CSI-RS or SS block. In the case of L1-RSRP measurements based on CSI-RS, the CSI-RS should be limited to single-port or dual-port CSI-RS. In the latter case, the reported L1-RSRP should be a linear average of the L1-RSRP measured on each port. The device can report measurements corresponding to up to four reference signals (CSI-RS or SS blocks), that is, up to four beams, in a single reporting instance. Each report is related to up to four reference signals or beams and includes the measured L1-RSRP for the strongest

[3] CSI-RSRP is defined as the linear average over the power contributions (in Watts) of the resource elements that carry CSI-RS configured for RSRP measurements within the identified measurement frequency region in the configured CSI-RS occasions. The CSI reference signals are transmitted on specific antenna ports. This measurement is applicable for connected mode only for both intra- and inter-frequency measurements.

beam and the difference between other beams' L1-RSRP measurements and the measured L1-RSRP of the best beam [14].

If a UE is configured with an *NZP-CSI-RS-ResourceSet* via RRC parameter *repetition* set to "on," it may assume that the CSI-RS resources within the *NZP-CSI-RS-ResourceSet* are transmitted with the same downlink spatial domain transmission filter, where the CSI-RS resources in the *NZP-CSI-RS-ResourceSet* are transmitted on different OFDM symbols. If *repetition* is set to "off," the CSI-RS resources within the *NZP-CSI-RS-ResourceSet* are transmitted with the same downlink spatial domain transmission filter. If the UE is configured with a *CSI-ReportConfig* and parameter *reportQuantity* is set to "cri-RSRP," or "none" and if the *CSI-ResourceConfig* for channel measurement (RRC parameter *resourcesForChannelMeasurement*) contains a *NZP-CSI-RS-ResourceSet* that is configured with the higher layer parameter *repetition* and without the higher layer parameter *trs-Info*, the UE can only be configured with the same number (1 or 2) of ports with the higher layer parameter *nrofPorts* for all CSI-RS resources within the set. If the UE is configured with the CSI-RS resource on the same OFDM symbol(s) as an SS/PBCH block, the CSI-RS and the SS/PBCH block are quasi-co-located with QCL TypeD, if applicable. Furthermore, the UE will not be configured with the CSI-RS in PRBs that overlap with those of the SS/PBCH block, and the same subcarrier spacing is used for both the CSI-RS and the SS/PBCH block [9].

4.1.2.4 Tracking Reference Signals

A UE must track and compensate time and frequency variations of its local oscillator in order to successfully receive downlink transmissions. The problem is exacerbated in higher radio frequencies. To assist the device in this task, a tracking reference signal can be configured. The TRS is not a CSI-RS, rather a TRS is a resource set consisting of multiple periodic NZP-CSI-RS. More specifically, a TRS consists of four single-port, density-3 CSI-RS located within two consecutive slots as shown in Fig. 4.7. The CRS-RS within the resource

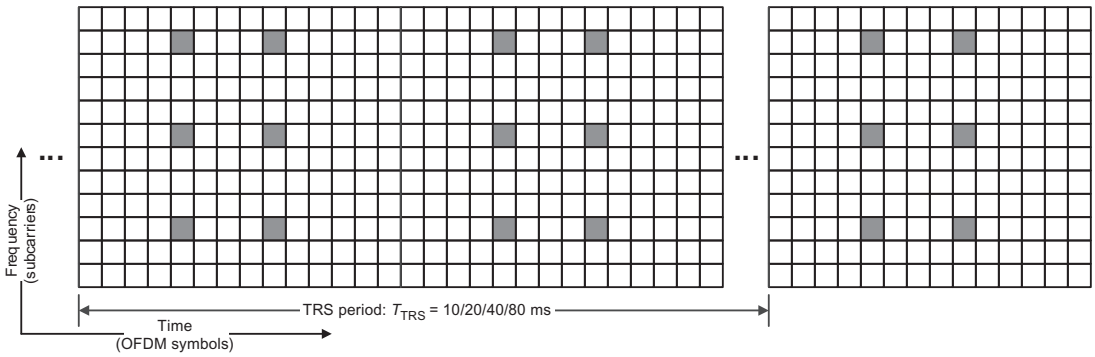


Figure 4.7
Example TRS structure (four single-port, Density-3 CSI-RS over two consecutive slots [9]).

set or the TRS can be configured with a periodicity of 10, 20, 40, or 80 ms. It must be noted that the exact set of time-frequency resource elements used for the TRS may vary; however, the two CSI-RS within a slot are always separated by four symbols in the time domain. This time-domain separation sets a limit for the maximum frequency error that can be compensated. Likewise, the frequency-domain separation of four subcarriers sets a limit for the maximum timing error that can be compensated. There is an alternative TRS structure with the same per-slot structure as the TRS structure shown in Fig. 4.7 with only two CSI-RS within a slot, compared to two consecutive slots for the TRS structure shown in the figure. In LTE, the CRS served the same purpose as the TRS; however, the TRS has relatively lower overhead, has one antenna port, and only present in two slots in every TRS period.

As we mentioned earlier, the CSI-RS for tracking or TRS is a special configuration of CSI-RS which is specifically configured for a UE. The TRS is used for fine time and frequency tracking as well as path delay spread and Doppler spread estimation. A UE in RRC_CONNECTED mode would receive information on a CSI-RS resource set which is configured specifically for the purpose of time/frequency tracking. In that case, the UE is configured with RRC parameter *trs-Info* and will assume that TRS is transmitted from the antenna port with the same port index of the configured NZP CSI-RS resources in the CSI-RS resource set. In frequency range 1, the UE may be configured with a CSI-RS resource set consisting of four periodic CSI-RS resources in two consecutive slots with two periodic CSI-RS resources in each slot, whereas in frequency range 2, the UE may be configured with a CSI-RS resource set of two periodic CSI-RS resources in one slot or with a CSI-RS resource set of four periodic CSI-RS resources in two consecutive slots with two periodic CSI-RS resources in each slot. The periodic CSI-RS resources in the CSI-RS resource set configured with RRC parameter *trs-Info* have the same periodicity, bandwidth and subcarrier location. The time-domain location of the TRS is determined by two periodic CSI-RS resources in a slot, or four periodic CSI-RS resources in two consecutive slots and via RRC signaling. The density of the TRS in one physical resource block is three REs per symbol. The TRS can be time-division multiplexed with the synchronization signal/PBCH block (SSB).

4.1.3 Control Channels

4.1.3.1 Physical Broadcast Channel

In order to select a PLMN and camp on a cell, a UE must perform a cell search in the supported frequency bands. The procedure requires the UE to achieve time and frequency synchronization with a specific cell. This enables decoding of PBCH which carries the MIB containing the critical system information necessary to decode transmissions on PDSCH. In NR, the SI is divided into minimum SI and other SI. The minimum SI is periodically broadcast and comprises basic information required for initial access and

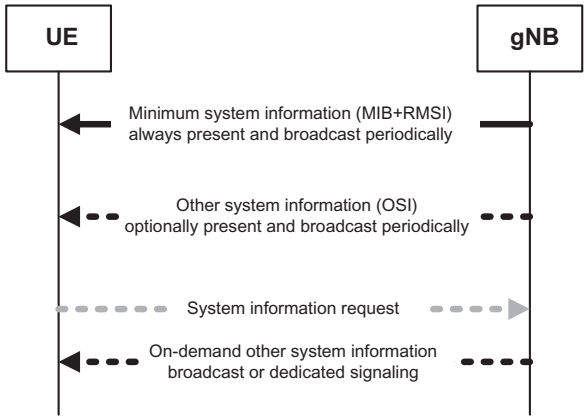


Figure 4.8
Transmission of system information in NR [11].

information for acquiring other SI broadcast periodically or scheduled on-demand. The other SI encompasses everything else not broadcast in the minimum SI message and may be either broadcast or individually transmitted to the UE. In the latter case, the on-demand transmission of other SI can be triggered by the network or based on a request from the UE. The change of SI can only occur at specific radio frames. Note that to ensure coverage and reliability, the SI may be transmitted a number of times with the same content within a modification period. For the minimum SI delivery, part of minimum SI is transmitted in PBCH. The remaining minimum SI (RMSI) is transmitted in the downlink shared channel. The initial BWP information is signaled by PBCH which contains the CORESET and PDSCH information for mapping the RMSI. Fig. 4.8 shows the process for transmission of various components of system information. Unlike LTE system where the minimum SI and a group of SI blocks were broadcast periodically, NR limits the amount of SI that is periodically broadcast and instead relies on less-frequent and on-demand transmission of the non-essential SI.

The MIB message (in ASN.1 format), which is carried in PBCH, consists of the following components [13]:

```
MIB ::= SEQUENCE {
            systemFrameNumber           BIT STRING (SIZE (6)),
              subCarrierSpacingCommon       ENUMERATED {scs15or60, scs30or120},
              ssb-SubcarrierOffset        INTEGER (0..15),
```

(*Continued*)

(Continued)

```
        dmrs-TypeA-Position          ENUMERATED {pos2, pos3},
        pdcch-ConfigSIB1
        cellBarred              ENUMERATED {barred, notBarred},
        intraFreqReselection          ENUMERATED {allowed, notAllowed},
        spareBIT STRING (SIZE (1))
}
PDCCH-ConfigSIB1 ::= SEQUENCE {
controlResourceSetZero ControlResourceSetZero,
searchSpaceZero SearchSpaceZero
}
```

In the MIB message which is mapped to BCH logical channel, *subCarrierSpacingCommon* parameter indicates the subcarrier spacing for SIB1, Msg2/4 for the initial access and the SI messages where values 15 and 30 kHz are applicable to sub-6 GHz and values 60 and 120 kHz are applicable to carrier frequencies above 6 GHz; *ssb-subcarrierOffset* is the frequency-domain offset between SSB and the overall resource block grid in number of subcarriers; *dmrs-TypeA-Position* indicates the position of the first downlink DM-RS; and *pdcch-ConfigSIB1* determines the bandwidth of PDCCH/SIB1 or the size of the CORESET containing common search space for PDCCH. In other words, the first field of *pdcch-ConfigSIB1* determines the common CORESET corresponding to the initial downlink BWP and the second field identifies the common search space of initial downlink BWP [13].

The RMSI is transmitted via the PDSCH by downlink assignment in an RMSI CORESET. The concept of CORESET was introduced in NR to identify a set of time-frequency resources consisting of multiple resource blocks in the frequency domain and one to three OFDM symbols in the time domain. The NR enables UE to be configured with multiple CORESETs, and each CORESET is associated with a UE-specific configured resource mapping scheme.

The PBCH payload size is 56 bits including 24-bit CRC. In NR, PBCH uses a single-antenna port transmission scheme, using the same antenna port as PSS and SSS within the same SS block. The periodicity of PBCH is 80 ms. The MIB data arrive at the PBCH processing unit in the form of one transport block (TB) every 80 ms and goes through the following steps as shown in Fig. 4.9: payload generation, scrambling, TB CRC calculation and attachment, channel coding, and rate matching. The coded and modulated bits of PBCH are

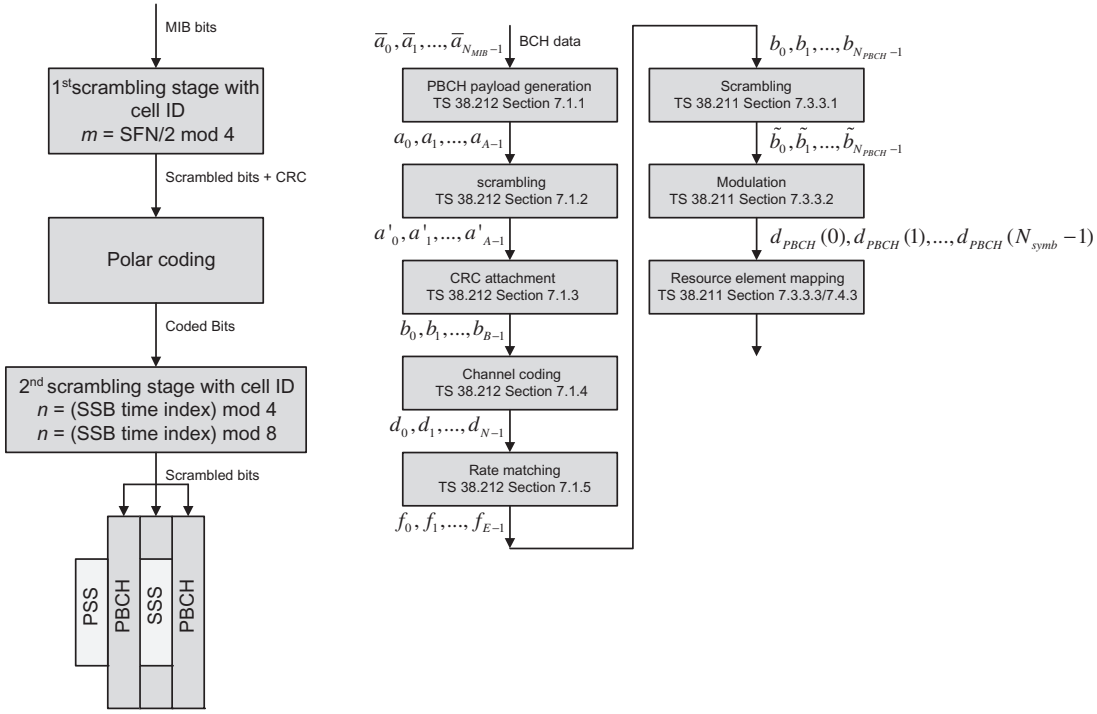


Figure 4.9
Physical layer processing of PBCH and mapping to time-frequency resources [5].

mapped onto the resource elements allocated for PBCH. As we will discuss later, the PBCH content is encoded using the polar code. Two scrambling operations are performed on PBCH which include one before CRC attachment and another one after the polar coding and rate matching. In the first scrambling stage, initialization based on Cell ID, the sequence is partitioned into four non-overlapping portions. The portion for transmission is selected based on the second and third least significant bits of the SFN. In the second scrambling stage, initialization based on Cell ID, the sequence is partitioned into four or eight non-overlapping portions. The portion for transmission is selected based on the second or third least significant bits of the SS block time index.

The physical layer processing of the PBCH is shown in Fig. 4.9. We denote the MIB bits in a TB delivered to the physical layer by $\overline{a}_0, \overline{a}_1, \ldots, \overline{a}_{N_{MIB}-1}$, where $N_{MIB} = 32$ bits is the payload size of the MIB. The lowest order information bit $\overline{a}_0$ is mapped to the most significant bit of the TB payload. Additional timing-related PBCH payload bits $\overline{a}_{N_{MIB}}, \overline{a}_{N_{MIB}+1}, \ldots, \overline{a}_{N_{MIB}+7}$ regenerated based on the least significant bits of the SFN, half frame, SS block index, and combined with the MIB payload (see Fig. 4.10). The $N_{MIB} + 8$ bits are interleaved according to an interleaving pattern specified in [7] prior to the first scrambling stage. In the first scrambling

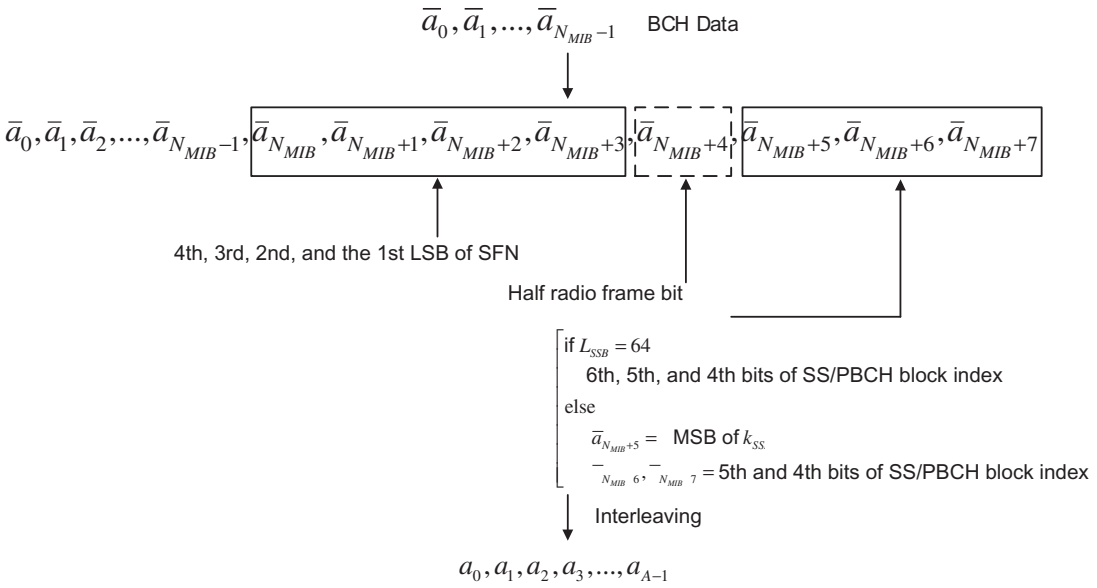


Figure 4.10
NR PBCH payload generation [7].

stage, the input bits to the scrambling unit a_i are scrambled according to $a'_i = (a_i + s_i) \bmod 2; \forall i = 0, 1, \ldots, N_{MIB} + 7$ where s_i is a sequence that is derived from a generic pseudo-random length-31 Gold sequence, which is initialized with $c_{init} = N_{ID}^{cell}$ at the start of each SFN whose value satisfies $SFN \bmod 8 = 0$. The s_i further depends on the half radio frame index, and the second and third least significant bits of the system frame number [7]. A CRC is calculated for the purpose of error detection on the entire BCH payload. The input bit sequence is denoted by $a'_0, a'_1, \ldots, a'_{A-1}$ and the parity bits by $p_0, p_1, \ldots, p_{L-1}$ where $A = N_{MIB} + 8$ is the payload size and $L = 24$ is the number of CRC parity bits. The parity bits are calculated and attached to the BCH payload using the generator polynomial $g_{CRC24C}(D) = D^{24} + D^{23} + D^{21} + D^{20} + D^{17} + D^{15} + D^{13} + D^{12} + D^8 + D^4 + D^2 + D + 1$, resulting in the sequence $b_0, b_1, \ldots, b_{B-1}$, where $B = A + L$. The information bits that are then delivered to the channel coding block are denoted by $c_0, c_1, \ldots, c_{K-1}$, where K is the number of input bits. They are encoded using a polar encoder by setting the parameters $n_{\max} = 9, I_{IL} = 1, n_{PC} = 0,$ and $n_{PC}^{wm} = 0$. Detailed PBCH channel encoding and decoding block diagram is shown in Fig. 4.11. The output of polar encoder is denoted by $d_0, d_1, \ldots, d_{N-1}$ where N is the number of coded bits. The input sequence to the rate matching function is $d_0, d_1, \ldots, d_{N-1}$ and the output bit sequence after rate matching is denoted by $f_0, f_1, \ldots, f_{E-1}$ where the rate matching output sequence length is $E = 864$ and the rate matching is performed by setting the parameter I_{BIL} to zero. The preceding polar coding and rate matching parameters will be explained in Section 4.1.7.1 [7].

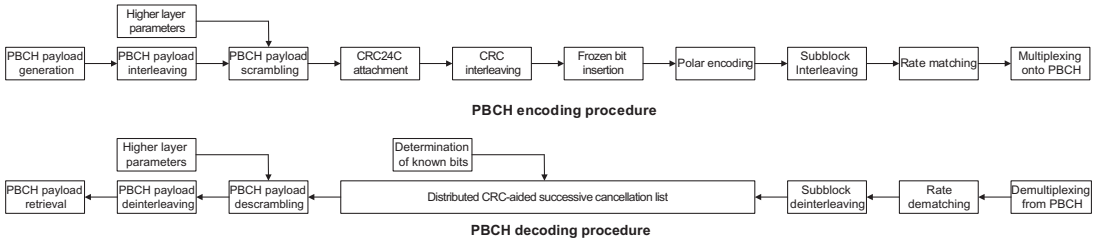


Figure 4.11
NR PBCH channel encoding/decoding block diagram [7,35].

The block of bits $b_0, b_1, \ldots, b_{N_{PBCH}-1}$, where N_{PBCH} denotes the number of bits transmitted on the PBCH, is scrambled prior to modulation, resulting in a block of scrambled bits $\tilde{b}_0, \tilde{b}_1, \ldots, \tilde{b}_{N_{PBCH}-1}$ in which $\tilde{b}_{(i)} = [b(i) + c(i + vN_{PBCH})] \bmod 2$ and $c(i)$ denotes a generic pseudo-random length-31 Gold sequence. The scrambling sequence is initialized with $c_{init} = N_{ID}^{cell}$ at the start of each SS/PBCH block. The parameter v is the two least significant bits of the SS/PBCH block index when $L_{\max} = 4$ and the three least significant bits of the SS/PBCH block index when $L_{\max} = 8$ or 64 where $L_{\max}$ denotes the maximum number of SS/PBCH blocks in an SS/PBCH period for a particular band (see Section 4.1.4.3). The block of scrambled bits $\tilde{b}_0, \tilde{b}_1, \ldots, \tilde{b}_{N_{PBCH}-1}$ is QPSK modulated, resulting in a block of complex-valued modulation symbols $d_{PBCH}(0), d_{PBCH}(1), \ldots, d_{PBCH}(N_{symb} - 1)$. The mapping of the modulated symbols to the physical resources is described in Section 4.1.4.3. The PBCH exploits a special type of DM-RS that is used for coherent detection and decoding of PBCH [6].

The total number of resource elements used for PBCH transmission per SS block is 576. Note that this number includes the resource elements for PBCH and the resource elements for the DM-RS needed for coherent demodulation of PBCH. Different numerologies can be used for SS/PBCH block transmission. However, to limit the need for devices to simultaneously search for SS/PBCH blocks of different numerologies, in many cases only a single SS block numerology is defined for a given frequency band. The DM-RS sequence $r_{PBCH}(m)$ for an SS/PBCH block is defined by $r_{PBCH}(m) = ([1 - 2c(2m)] + j[1 - 2c(2m + 1)]/\sqrt{2})$ where $c(n)$ is a length-31 Gold sequence generated by the pseudo-random sequence generator defined in [6] and initialized at the start of each SS/PBCH block with $c_{init} = 2^{11}(\bar{i}_{SSB} + 1)\left(\lfloor N_{ID}^{cell}/4 \rfloor + 1\right) + 2^6(\bar{i}_{SSB} + 1) + (N_{ID}^{cell} \bmod 4)$ [6]. If $L = 4$, $\bar{i}_{\mathrm{SSB}} = i_{\mathrm{SSB}} + 4n_{hf}$ where n_{hf} denotes the number of the half-frame in which PBCH is transmitted in a frame with $n_{hf} = 0$ for the first half-frame in the frame and $n_{hf} = 1$ for the second half-frame in the frame, and i_{SSB} is the two least significant bits of the SS/PBCH block index. In the case that $L = 8$ or $L = 64$, $n_{hf} = 0$, and $\bar{i}_{\mathrm{SSB}} = i_{\mathrm{SSB}}$ are the three least significant bits of the SS/PBCH block index. Note that L denotes the maximum number of SS/PBCH block beams in an SS/PBCH block period for a particular band [3].

The sequence of complex-valued QPSK-modulated symbols $d_{PBCH}(0)$, $d_{PBCH}(1), \ldots, d_{PBCH}(N_{symb} - 1)$ containing the PBCH information are scaled by a factor β_{PBCH} to adjust the PBCH transmit power and then mapped in sequence starting with $d_{PBCH}(0)$ to resource elements (k, l) provided that they are not used for PBCH DM-RSs (see Fig. 4.12). The sequence of complex-valued symbols $r_{PBCH}(0), r_{PBCH}(1), \ldots, r_{PBCH}(143)$ containing the DM-RSs for the SS/PBCH block is scaled by a factor of $\beta_{PBCH}^{DM\text{-}RS}$ in order to

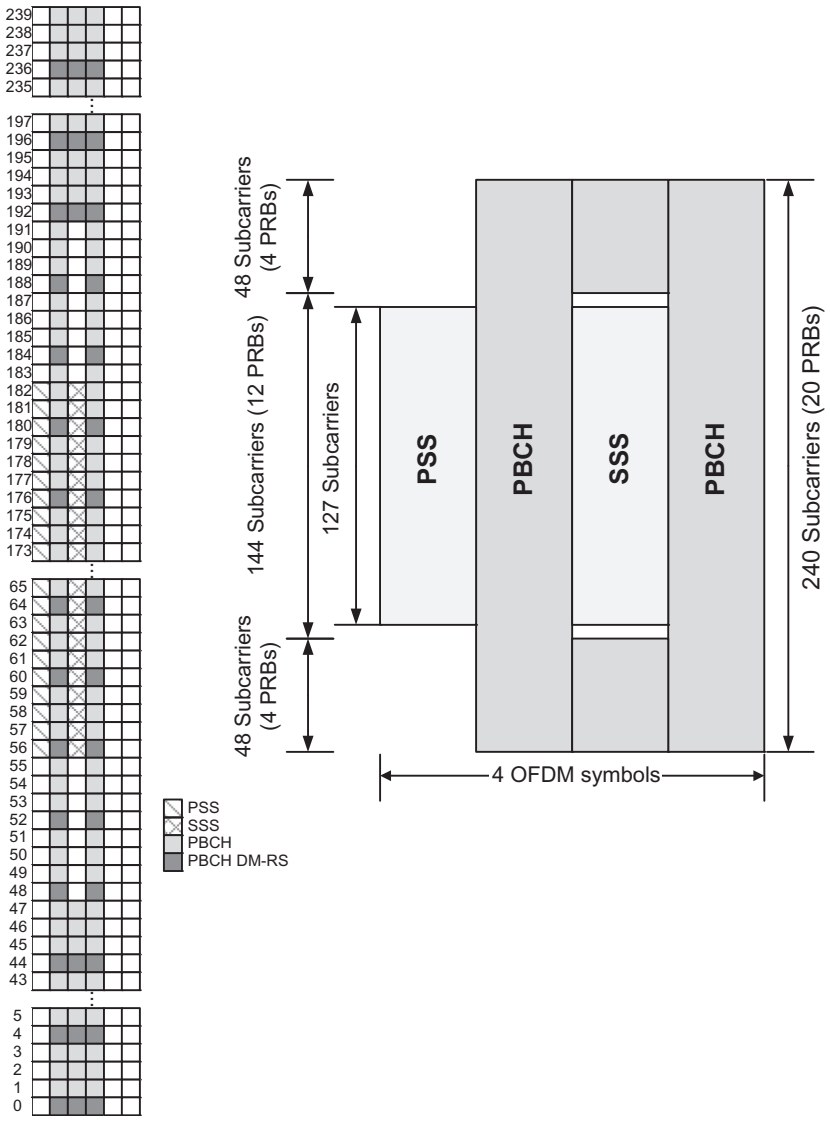


Figure 4.12
Structure of PBCH in time/frequency domain [30].

Table 4.3: Resource mapping within an SS/PBCH block for PSS, SSS, PBCH, and DM-RS for PBCH [6].

Physical Channel/Signal	OFDM Symbol Number l	Subcarrier Number k
	Relative to the Start of an SS/PBCH Block	Relative to the Start of an SS/PBCH Block
PSS	0	56,57,. . .,182
SSS	2	56,57,. . .,182
Null	0	0,1,. . .,55,183,184,. . .,239
	2	48,49,. . .,55,183,184,. . .,191
PBCH	1,3	0,1,. . .,239
	2	0,1,. . .,47, 192,193,. . .,239
DM-RS for PBCH	1,3	$0+v, 4+v, 8+v, \ldots, 236+v$
	2	$0+v, 4+v, 8+v, \ldots, 44+v$ $192+v, 196+v, \ldots, 236+v$

adjust the PBCH DM-RS transmit power. They are then mapped to resource elements (k, l) in increasing order of first k (frequency index) and then l (time index), within one SS/PBCH block. As shown in Table 4.3, the location of PBCH DM-RS is dependent on parameter $v = N_{ID}^{cell} \bmod 4$ and is shifted in frequency with different N_{ID}^{cell} values (see Fig. 4.13 for an example).

4.1.3.2 Physical Downlink Control Channel

The data transmission in NR in downlink/uplink direction is generally controlled via MAC scheduling. Each device monitors a number of PDCCHs, typically once per slot, although it is possible to configure more frequent monitoring to support traffic requiring very low latency. Upon detection of a valid PDCCH, the device follows the scheduling decision and receives (or transmits) one unit of data, known as a transport block. The PDCCHs are transmitted in one or more CORESETs each of length one to three OFDM symbol(s). Unlike LTE, where control channels span the entire carrier bandwidth, the bandwidth of a CORESET can be configured. In NR, a flexible slot format can be configured for a UE by cell-specific and/or UE-specific higher layer signaling in a semi-static downlink/uplink assignment manner, or by dynamically signaling via DCI in group-common PDCCH (GC-PDCCH). When the dynamic signaling is configured, a UE should monitor GC-PDCCH which carries dynamic slot format indication (SFI). When a device enters the connected state, it has already obtained the information from PBCH about the CORESET where it can find the control channel used to schedule the RMSI. The CORESET configuration obtained from PBCH also defines and activates the initial bandwidth part in the downlink. The initial active uplink bandwidth part is obtained from the SI scheduled using the downlink PDCCH.

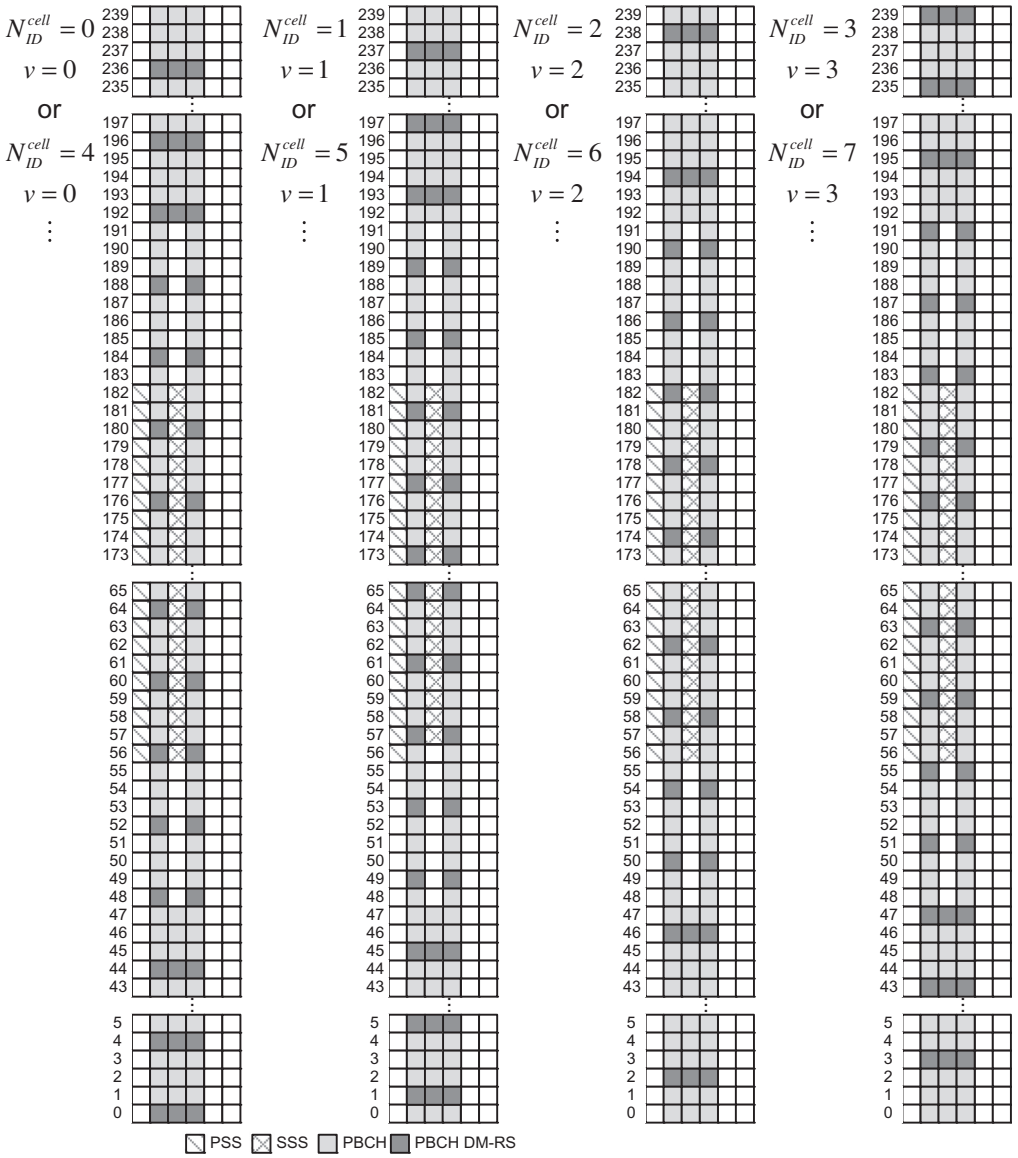


Figure 4.13
PBCH DM-RS location shift by physical cell ID [30].

In NR, the device typically attempts to blindly decode candidate PDCCHs transmitted from the network using one or more search spaces. However, there are some differences compared to LTE based on the different design targets for NR as well as the experience learned from LTE deployments. Unlike LTE PDCCH, the PDCCH in NR does not span the entire carrier bandwidth. This is due to the fact that NR devices may not be able to operate over

the full carrier bandwidth of the gNB. The PDCCH in NR is designed to support UE-specific beamforming, which is the result of beam-centric physical-layer design of NR and a requirement when operating in mmWave bands with challenging link budgets.

4.1.3.2.1 Structure and Physical Layer Processing of PDCCH

A UE-specific PDCCH is used to schedule downlink and uplink transmissions on PDSCH and PUSCH, respectively. The DCI on PDCCH contains downlink assignments including modulation and coding format, resource allocation, and HARQ information related to DL-SCH; uplink scheduling grants including modulation and coding format, resource allocation, and HARQ information related to UL-SCH. The control channels are formed by aggregation of control channel elements (CCE), each control channel element consisting of a set of resource element groups (REG). Different code rates for the control channels are realized by aggregating different number of control channel elements. Polar code is used for PDCCH channel coding. Each resource element group carrying PDCCH includes its own DM-RS. QPSK modulation is used for PDCCH modulation.

A resource element is the smallest unit of the resource grid consisting of one subcarrier in frequency domain and one OFDM symbol in time domain. A PDCCH corresponds to a set of resource elements carrying DCI. Each NR control channel element consists of six REGs where a REG is equivalent to one resource block (12 resource elements in the frequency domain) over one OFDM symbol. The CCE size is designed such that at least one UE-specific DCI can be transmitted within one CCE with lower code rates. An NR REG bundle is further defined comprising 2, 3, or 6 REGs, which provides: (1) it determines the precoder cycling granularity (which affects the channel estimation performance) and (2) it is the interleaving unit for the distributed REG mapping. An NR PDCCH candidate consists of a set of CCEs, that is, 1, 2, 4, 8, or 16, corresponding to aggregation levels (ALs) 1, 2, 4, 8, 16, respectively. A control search space consists of a set of PDCCH candidates and is closely associated with the ALs, the number of decoding candidates for each AL, and the set of CCEs for each decoding candidate. A search spacein NR Rel-15 is associated with a single CORESET.

As shown in Fig. 4.14 CORESET is defined as a set of REGs with a given numerology. In the frequency domain, a CORESET is defined as a set of contiguous or distributed physical resource blocks configured using a six-PRB granularity, within which the UE attempts to blindly decode the DCI. There is no restriction on the maximum number of segments for a given CORESET. In the time domain, a CORESET spans 1, 2, or 3 contiguous OFDM symbol, and the exact duration is signaled to the UE via broadcast SI or UE-specific RRC signaling depending on whether it is a common CORESET or UE-specific CORESET. Compared to LTE PDCCH, the configurability of the CORESETs enable efficient resource sharing between downlink control and shared channels, thereby allowing efficient layer-1 signaling overhead management. One of the factors which could impact

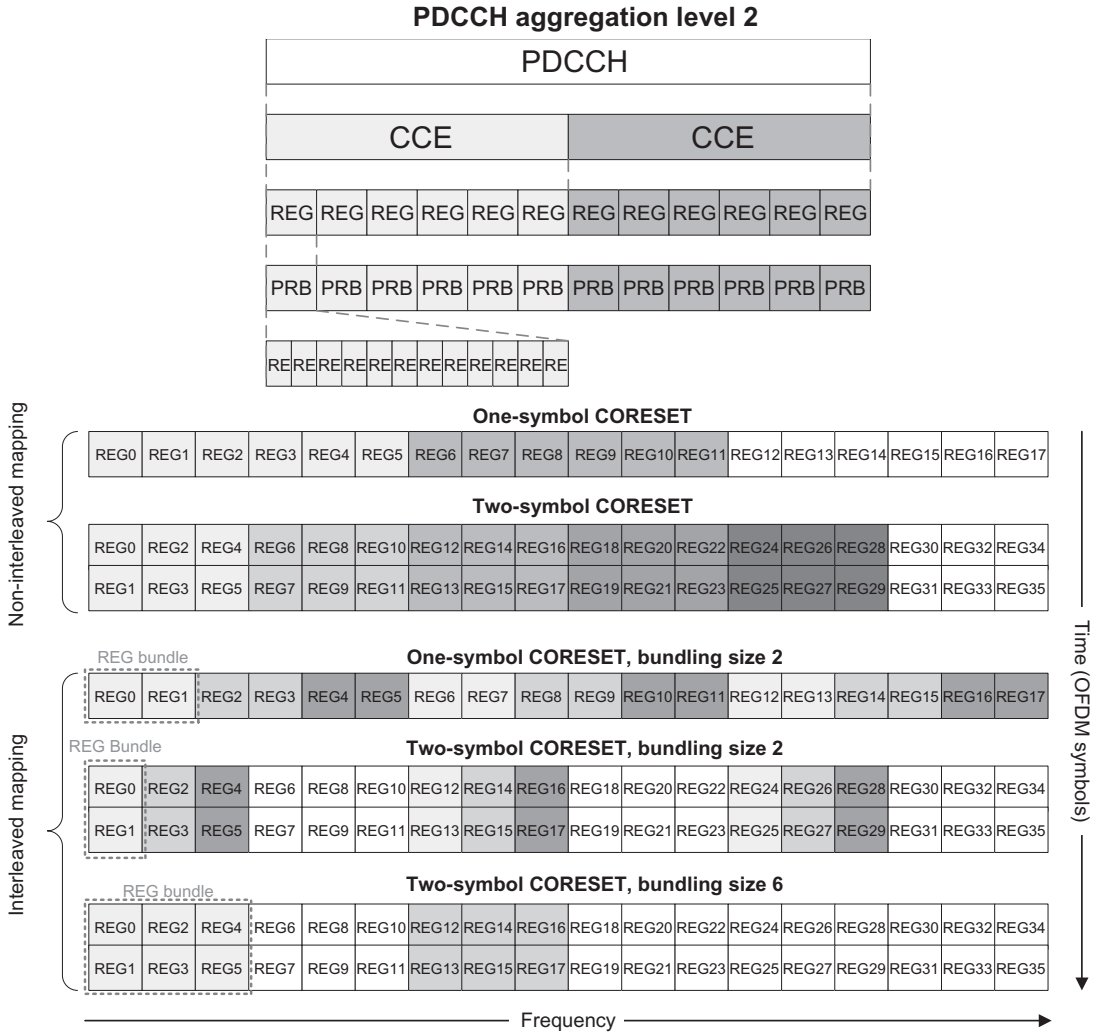


Figure 4.14
Illustration of RE, REG, CCE, REG bundle, and CORESET and example mappings of CCE to REG Bundles [14,68].

the time-domain duration of a CORESET is the bandwidth of the corresponding carrier, where more control symbols may be allowed for smaller bandwidths. For example, assuming a CORESET consists of 48 PRBs with 2 OFDM symbols, there are 16 CCEs that could accommodate up to 2 PDCCH candidates at AL-8 or a single candidate at AL-16. Furthermore, there can be multiple CORESETs inside the system bandwidth; thus the CORESET may not fully occupy the system bandwidth in the frequency domain. Downlink power adjustment can be applied to CORESETs that occupy narrower frequency regions within the carrier bandwidth, depending on desired coverage and link

budget. In that case, one or two OFDM symbols may not be sufficient. One-symbol CORESET offers benefits from the perspective of latency and control overhead adjustments especially when there are few UEs in the cell or when the coverage target is limited (e.g., small cell deployments). The maximum CORESET duration that may be configured in a cell is implicitly signaled via PBCH. A UE may be configured with one or more CORESETs (using UE-specific or common higher layer signaling) with a maximum of three CORESETs per configured (downlink) BWP. Limiting the maximum number of CORESETs is beneficial for enabling more practical RRC signaling and better UE dimensioning. Note that the scheduling flexibility may not be impacted by limiting the maximum number of CORESETs since different monitoring occasions can be flexibly configured associated with the same CORESET. It is important to note that the concept of PDCCH monitoring periodicity is defined per search space set and is not configured at the CORESET level. Every configured search space with a certain monitoring periodicity (in terms of slots and starting symbols within the monitored slots) is associated with a CORESET. For a CORESET configured by UE-specific RRC signaling, some of the configured parameters include frequency-domain resources, starting OFDM symbol, CORESET duration, REG bundle size, transmission type (i.e., interleaved or non-interleaved), and precoding assumptions for channel estimation filtering [53].

As we mentioned earlier, a PDCCH consists of one or more CCEs. A CORESET consists of $N_{RB}^{CORESET}$ resource blocks in the frequency domain, determined by the RRC parameter *frequencyDomainResources* in *ControlResourceSet* information element, and $N_{symb}^{CORESET} \in \{1, 2, 3\}$ OFDM symbols in the time domain, defined by the RRC parameter *duration* in the *ControlResourceSet* information element, where $N_{symb}^{CORESET} = 3$ is supported, if the RRC parameter *dmrs-TypeA-Position* is set to 3 [6]. A control channel element consists of six REGs where a REG is equivalent to one resource block over one OFDM symbol. The REGs within a CORESET are numbered in increasing order in a time-first manner, starting with 0 for the first OFDM symbol and the lowest numbered resource block in the CORESET. A UE can be configured with multiple CORESETs. Each CORESET is associated with only one CCE-to-REG mapping (see Fig. 4.14). There is a direct correspondence between the number of CCEs and the AL, for example, for ALs 1, 2, 4, 8, and 16, there will be 1, 2, 4, 8, and 16 CCEs, respectively [6]. The time-frequency structure of REG, CCE, REG bundle, and CORESET as well as example mappings of CCE to REG bundles are illustrated Fig. 4.14.

The PDCCH processing steps are illustrated in Fig. 4.15. At a high level, the PDCCH processing in NR is similar to that of LTE ePDCCH than LTE PDCCH in the sense that each PDCCH is processed independently. As shown in the figure, the entire DCI bits are used to calculate the CRC parity bits. Let us assume that $a_0, a_1, \ldots, a_{N_{DCI}-1}$ denote the DCI input bits, and $p_0, p_1, \ldots, p_{L-1}$ represent the parity bits, where N_{DCI} and $L = 24$ are the payload size and the number of parity bits, respectively. Let us assume that $a'_0, a'_1, \ldots, a'_{N_{DCI}-1}$ is bit

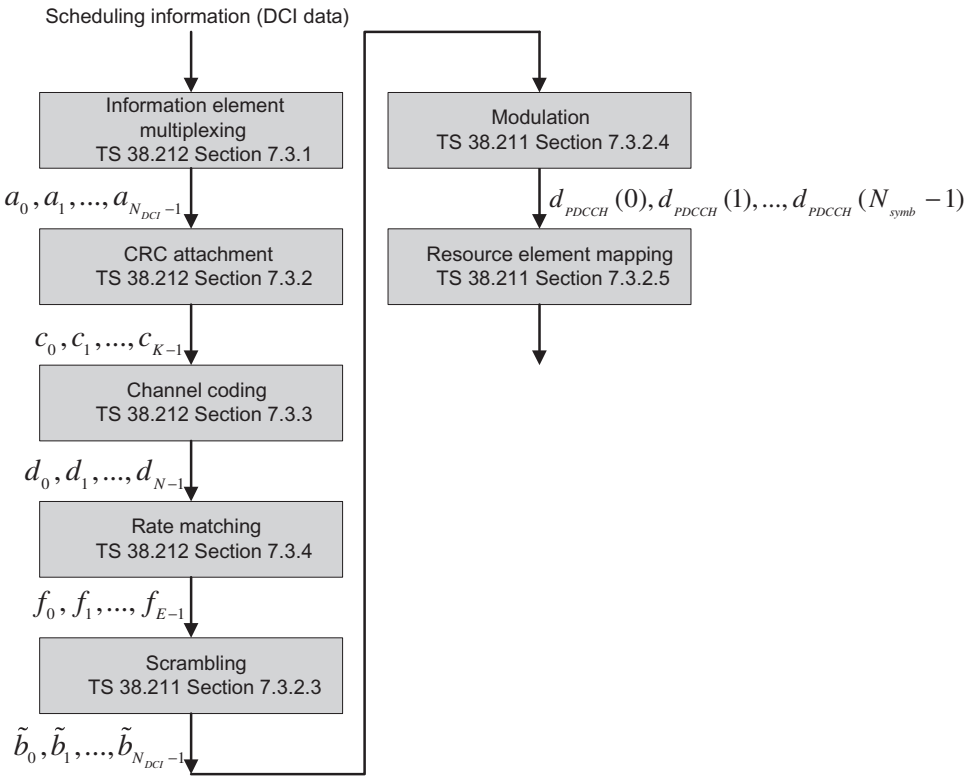


Figure 4.15
Physical layer processing of NR PDCCH [6,7].

sequence such that $a'_i = 1\ \forall i = 0, 1, \ldots, L-1$ and $a'_i = a'_{i-L}\ \forall i = L, L+1, \ldots, L+N_{DCI}-1$. The parity bits are computed with input bit sequence $a'_0, a'_1, \ldots, a'_{N_{DCI}+L-1}$ using the generator polynomial $g_{CRC24C}(D) = D^{24} + D^{23} + D^{21} + D^{20} + D^{17} + D^{15} + D^{13} + D^{12} + D^8 + D^4 + D^2 + D + 1$. The output bit sequence is given as $b_0, b_1, \ldots, b_{B-1}$ where $b_k = a_k \forall k = 0, 1, \ldots, N_{DCI} - 1$ and $b_k = p_{k-N_{DCI}} \forall k = N_{DCI}, N_{DCI+1}, \ldots, N_{DCI+L-1}$. Following the attachment of the CRC bits, the sequence is scrambled with the corresponding 16-bit RNTI $x_{RNTI}(0), x_{RNTI}(1), \ldots, x_{RNTI}(15)$, where $x_{RNTI}(0)$ corresponds to the MSB of the RNTI binary value, resulting in the sequence of bits $c_0, c_1, \ldots, c_{K-1}$ where $c_k = b_k \forall k = 0, 1, \ldots, N_{DCI} + 7$ and $c_k = [b_k + x_{RNTI}(k - N_{DCI} - 8)] \bmod 2 \forall k = N_{DCI} + 8, N_{DCI} + 9, \ldots, N_{DCI} + 23$ [6,7].

The PDCCH encoding stages are shown in Fig. 4.15. The K scrambled information bits are delivered to the channel coding block and are polar coded, by setting the encoder parameters to the following $n_{\max} = 9, I_{IL} = 1, n_{PC} = 0$, and $n_{PC}^{wm} = 0$. The encoding process produces N bits which are denoted as $d_0, d_1, \ldots, d_{N-1}$. The rate matching for polar coded bits is performed on per coded block and consists of subblock interleaving, bit collection, and bit interleaving. Detailed PDCCH channel encoding and decoding block diagram is shown in Fig. 4.16.

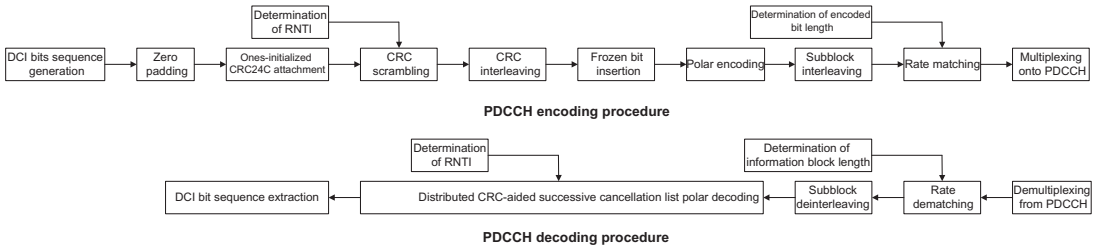


Figure 4.16
NR PDCCH channel encoding/decoding block diagram [7,35].

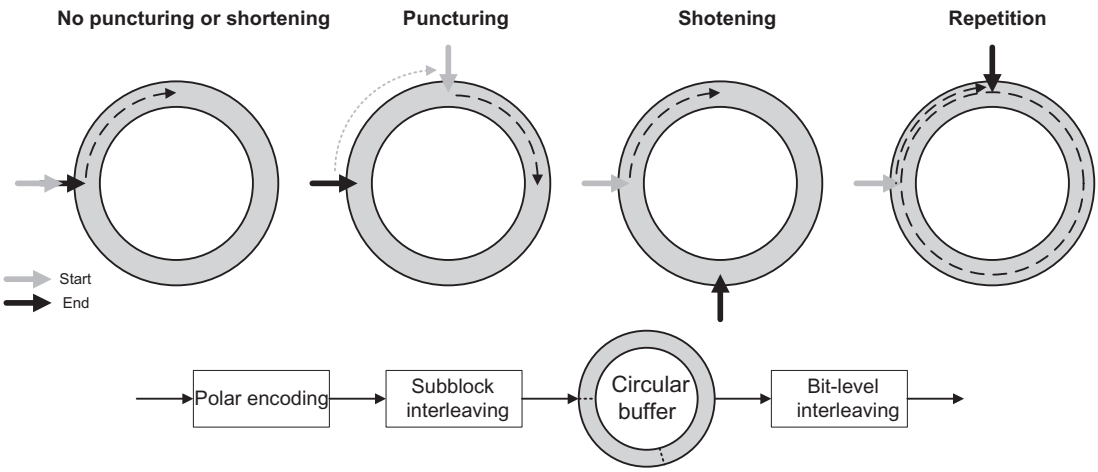


Figure 4.17
Encoding of NR PDCCH and rate matching variants [68].

The input bit sequence to rate matching function is denoted by $d_0, d_1, \ldots, d_{N-1}$ and the output is denoted by $f_0, f_1, \ldots, f_{E-1}$. The input bits to the subblock interleaver are divided into 32 subblocks and the output bits are generated according to $y_n = d_{J(n)}$ where $J(n) = NP\left(\lfloor 32n/N \rfloor\right)/32 + \left(n \bmod N/32\right) \forall n = 0, 1, \ldots, N-1$ and the subblock interleaver pattern is defined in [7]. The repetition, puncturing, or shortening of polar code is performed in the following manner: $N = 2^n$ coded bits at the output of polar encoder is written into a length-N circular buffer in an order that is predefined for a given value of N. As shown in Fig. 4.17, to obtain M coded bits for transmission, puncturing is realized by selecting bits from position $N - M$ to position $N - 1$ from the circular buffer, shortening is realized by selecting bits from position 0 to position $M - 1$ from the circular buffer, and repetition is realized by selecting all bits from the circular buffer, and additionally repeating $M - N$ consecutive bits from the circular buffer [7].

For each CORESET, there is an associated CCE-to-REG mapping based on the REG bundle (see Fig. 4.14). A REG bundle is a set of REGs across which the device can assume the

precoding is constant. This property can be exploited to improve the channel estimation performance, which is similar to PRB bundling for PDSCH. The CCE-to-REG mapping can be either interleaved or non-interleaved, depending on the characteristics of the transmission channel, that is, frequency-flat or frequency-selective fading channel. There is only one CCE-to-REG mapping for a given CORESET; however, since the mapping is a property of the CORESET, multiple CORESETs can be configured with different mappings. The CCE-to-REG mapping for a CORESET can be interleaved or non-interleaved, configured by the RRC parameter *cce-REG-MappingType* in the *ControlResourceSet* information element and is described by REG bundles. The REG bundle i is defined as REGs $\{iL_{REG}, iL_{REG}+1, \ldots, iL_{REG}+L_{REG}-1\} \forall i = 0, 1, \ldots, N_{REG}^{CORESET}/L_{REG}-1$ where L_{REG} is the size of the REG bundle and $N_{REG}^{CORESET} = N_{RB}^{CORESET} N_{symb}^{CORESET}$ is the number of REGs in the CORESET. The jth CCE consists of REG bundles $\{\Phi(6j/L_{REG}), \Phi(6j/L_{REG}+1), \ldots, \Phi(6j/L_{REG}+6/L_{REG}-1)\}$, where $\Phi(\cdot)$ denotes an interleaving function. In case of non-interleaved CCE-to-REG mapping $L_{REG} = 6$ and $\Phi(j) = j$, whereas in the case of interleaved CCE-to-REG mapping $L_{REG} = \{2, 6\}$ for $N_{symb}^{CORESET} = 1$ and $L_{REG} \in \left\{N_{symb}^{CORESET}, 6\right\}$ for $N_{symb}^{CORESET} \in \{2, 3\}$ where L_{REG} is configured by the RRC parameter *reg-BundleSize*. The interleaving function is defined by $\Phi(j) = rC + c + n_{shift} \mathrm{mod} N_{REG}^{CORESET}/L_{REG}; j = cR + r; r = 0, 1, \ldots, R-1; c = 0, 1, \ldots, C-1$ and $C = N_{REG}^{CORESET}/L_{REG}R$ where $R \in \{2, 3, 6\}$ is given by the higher layer parameter *interleaverSize*. Other parameters are defined as follows: $n_{shift} = N_{ID}^{cell}$ for a PDCCH transmitted in a CORESET configured by the PBCH or SIB1, and $n_{shift} \in \{0, 1, \ldots, 274\}$ is given by the RRC parameter *shiftIndex*. The UE is not expected to monitor configurations for which C is not an integer. For both interleaved and non-interleaved mappings, the same precoding is used within an REG bundle, if the higher layer parameter *precoderGranularity* equals L_{REG}. The same precoding is used across all REGs within the set of contiguous resource blocks in the CORESET, if the higher layer parameter *precoderGranularity* equals the size of the CORESET in the frequency domain. For a CORESET configured by PBCH, $L_{REG} = 6, R = 2$ and the same precoding is used within the REG bundle [6]. Unlike LTE, where the length of the control region can vary dynamically as indicated by PCFICH, a CORESET in NR is of fixed size. This is important from an implementation perspective, both for the UE and the network. From the UE perspective, a pipelined implementation is simpler, if the device can directly start to decode PDCCH without having to first decode another control channel. Various REG-to-CCE mapping options are shown in Fig. 4.18.

During the PDCCH detection and decoding process, the UE needs to estimate the channel using the reference signals associated with the PDCCH candidate being decoded. A single antenna port is used for PDCCH transmission which means any transmit diversity or multi-user MIMO scheme is handled in a device-transparent manner. The PDCCH has its own DM-RS, based on the same pseudo-random sequence that is used for PDSCH, that is, the pseudo-random sequence is generated across all the common resource blocks in the

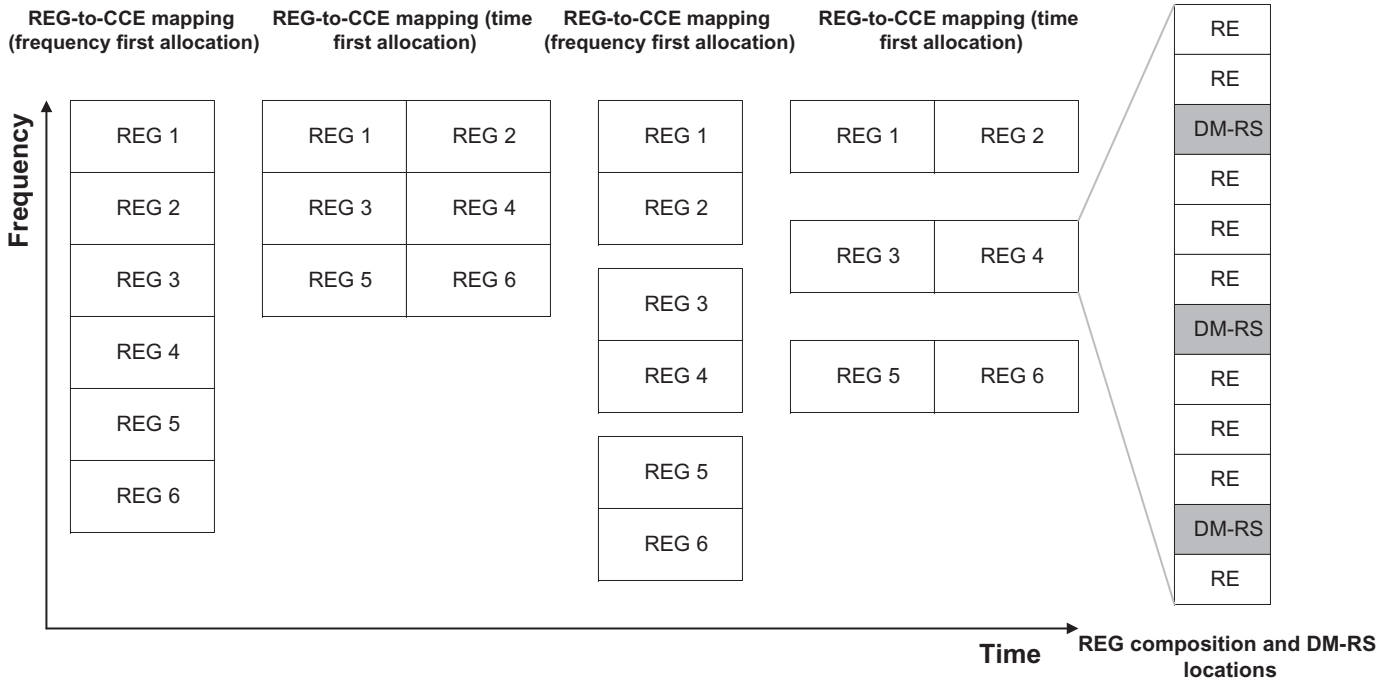


Figure 4.18
REG-to-CCE mapping options and REG composition [6].

frequency domain, but transmitted only in the resource blocks used for PDCCH (with one exception as discussed below). However, during initial access, the location of the common resource blocks is not known to UE as this information is signaled as part of the minimum SI. Therefore, for CORESET 0 configured by PBCH, the sequence is generated starting from the first resource block in the CORESET. The RRC parameters that define a CORESET are as follows [13]:

```
ControlResourceSet:: =        SEQUENCE {
     controlResourceSetId                 ControlResourceSetId,
     frequencyDomainResources              BIT STRING (SIZE (45)),
     duration                             INTEGER (1..maxCoReSetDuration),
//maxCoReSetDuration = 3
     cce-REG-MappingType CHOICE {
               interleaved                   SEQUENCE {
                         reg-BundleSize               ENUMERATED {n2, n3, n6},
                         interleaverSize               ENUMERATED {n2, n3, n6},
                         shiftIndex                                    INTEGER(0..
maxNrofPhysicalResourceBlocks-1)
               },
               nonInterleaved         NULL
     },
     precoderGranularity ENUMERATED {sameAsREG-bundle, allContiguousRBs},
     tci-StatesPDCCH       SEQUENCE(SIZE (1..maxNrofTCI-StatesPDCCH)) OF TCI-StateId
     tci-PresentInDCI       ENUMERATED {enabled}       OPTIONAL
     pdcch-DMRS-ScramblingID   BIT STRING (SIZE (16))       OPTIONAL
}
```

In the preceding definition [6,9],

- *controlResourceSetId* corresponds to L1 parameter *CORESET-ID* whose value 0 identifies the common CORESET configured in MIB and in *ServingCellConfigCommon* and values $1, 2, \ldots,$ *maxNrofControlResourceSets*-1 identify the CORESETs configured by dedicated signaling. The *controlResourceSetId* is unique among the BWPs of a serving cell.
- *frequencyDomainResources* corresponds to the L1 parameter *CORESET-freq-dom*. Each bit corresponds a group of six RBs, with the grouping start from PRB 0, which is fully

contained in the bandwidth part within which the CORESET is configured. The most significant bit corresponds to the group of the lowest frequency which is fully contained in the bandwidth part within which the CORESET is configured, each subsequent lower significant bit corresponds to the next lowest frequency group that are fully contained within the bandwidth part in which the CORESET is configured. The bits corresponding to a group not fully contained within the bandwidth part in which the CORESET is configured are set to zero.

- *duration* is the contiguous time duration of the CORESET in number of symbols.
- *cce-reg-MappingType* identifies the mapping method of CCE-to-REG.
- *reg-BundleSize* is the number of REGs within an REG bundle corresponding to L1 parameter *CORESET-REG-bundle-size*.
- interleaveSize corresponds to L1 parameter CORESET-interleaver-size.
- shiftIndex corresponds to CORESET-shift-index.
- *precoderGranularity* denotes the precoder granularity in frequency domain. It corresponds to L1 parameter *CORESET-precoder-granularity*.
- *tci-StatesPDCCH* is a reference to a configured transmission configuration indication (TCI)[4] state providing QCL configuration/indication for PDCCH.

[4] Downlink beamforming is typically transparent to the UE, that is, the device does not need to know which beam is used at the transmitter. However, NR also supports beam indication, which implies that the UE is informed of a certain PDSCH and/or PDCCH transmission using the same transmit beam as a configured reference signal (CSI-RS or SS block). The beam indication is based on the (downlink signaling of) transmission configuration indication (TCI) states. Each TCI state includes information about a reference signal, for example, a CSI-RS or an SS block. By associating a certain downlink transmission (PDCCH or PDSCH) with a certain TCI, the network informs the UE that it can assume the upcoming downlink transmission uses the same spatial filter as the reference signal associated with that TCI. A device can be configured with up to 64 candidate TCI states. The beam indication for PDCCH is done by assigning a subset of the *M* configured candidate states via RRC signaling to each configured CORESET. Using MAC signaling, the network can dynamically indicate a specific TCI state, within the per-CORESET-configured subset, to be valid. When monitoring PDCCH within a certain CORESET, the device can assume that the PDCCH transmission uses the same spatial filter as the reference signal associated with the MAC-indicated TCI. In other words, if the device has determined a suitable receiver-side beam direction for reception of the reference signal, it can assume that the same beam direction is suitable for reception of the PDCCH. For PDSCH beam indication, there are two alternatives depending on the scheduling offset, that is, depending on the transmission timing of PDSCH relative to the corresponding PDCCH carrying scheduling information for that PDSCH. If this scheduling offset is larger than *N* symbols, the DCI of the scheduling assignment may explicitly indicate the TCI state for the PDSCH transmission. To enable this, the device is initially configured with a set of up to eight TCI states from the originally configured set of candidate TCI states. A three-bit indicator within the DCI then indicates the exact TCI state which is valid for the scheduled PDSCH transmission. If the scheduling offset is smaller or equal to *N* symbols, the device should instead assume that the PDSCH transmission is QCL with the corresponding PDCCH transmission. In other words, the TCI state for the PDCCH state indicated by MAC signaling should be assumed to be valid for the corresponding scheduled PDSCH transmission. The reason for limiting the dynamic TCI selection based on DCI signaling to the scenarios where the scheduling offset is larger than a certain value is that for shorter scheduling offsets, there will not be sufficient time for the UE to successfully decode the TCI information within the DCI and to adjust the receive beam accordingly before the PDSCH is received [14].

- tci-PresentInDCI corresponds to L1 parameter CORESET-precoder-granularity.
- *pdcch-DMRS-ScramblingID* is the PDCCH DM-RS scrambling initialization.

The DM-RSs associated with a given PDCCH candidate are mapped to every fourth subcarrier in a REG, that is, the reference signal overhead is one-fourth. This is a denser reference signal pattern relative to LTE, which has a reference signal overhead of one-sixth; however, an LTE device can interpolate channel estimates across time and frequency as a result of cell-specific reference signals common to all devices and present regardless of control channel transmission. The use of a dedicated reference signal per PDCCH candidate is advantageous, despite the slightly higher overhead, since it allows different type of device-transparent beamforming schemes. By using a beamformed control channel, the coverage and performance can be enhanced compared to the non-beamformed control channels in LTE. This is an essential part of the beam-centric design of NR [14].

When attempting to decode a PDCCH candidate occupying certain number of CCEs, the device can compute the REG bundles that constitute the PDCCH candidate. Channel estimation must be performed per REG bundle as the network may change precoding across REG bundles. In general, this results in sufficiently accurate channel estimates for PDCCH detection. However, it is also possible to configure the device to assume the same precoding across contiguous resource blocks in a CORESET, thereby allowing the device to perform frequency-domain interpolation of the channel estimates. This also implies that the device may use reference signals referred to as wideband reference signals outside the PDCCH region that it is trying to detect (see Fig. 4.19). The QCL concept is also applicable to the reference signals. If the UE has a priori knowledge about the QCL of two reference signals, it can exploit this property to improve the channel estimation and to manage different receive beams at the device. If no QCL is configured for a CORESET, the UE assumes that PDCCH candidates are quasi-co-located with the SS/PBCH block with respect to delay spread, Doppler spread, Doppler shift, average delay, and spatial RX parameters. This is a reasonable assumption as the device has been able to receive and decode the PBCH in order to access the system.

The block of bits $b(0), b(1), \ldots, b(N_{PDCCH} - 1)$, where N_{PDCCH} denotes the number of bits transmitted on PDCCH, is scrambled prior to modulation, resulting in a block of scrambled bits $\tilde{b}(0), \tilde{b}(1), \ldots, \tilde{b}(N_{PDCCH} - 1)$ where $\tilde{b}(i) = [b(i) + c(i)] \bmod 2$, in which $c(i)$ is a length-31 Gold sequence generated by the pseudo-random sequence generator defined in [6] and initialized with $c_{init} = \left(n_{RNTI} 2^{16} + n_{ID}\right) \bmod 2^{31}$. For a UE-specific search space, $n_{ID} \in \{0, 1, \ldots, 65535\}$ is set by the RRC parameter *pdcch-DMRS-ScramblingID;* otherwise, $N_{ID} = N_{ID}^{cell}$ and n_{RNTI} is determined by the C-RNTI for a PDCCH in a UE-specific search space, when the RRC parameter *pdcch-DMRS-ScramblingID* is configured; otherwise $n_{RNTI} = 0$. The block of scrambled bits $\tilde{b}(0), \tilde{b}(1), \ldots, \tilde{b}(N_{PDCCH} - 1)$ is then QPSK modulated, resulting in a block of complex-valued modulation symbols

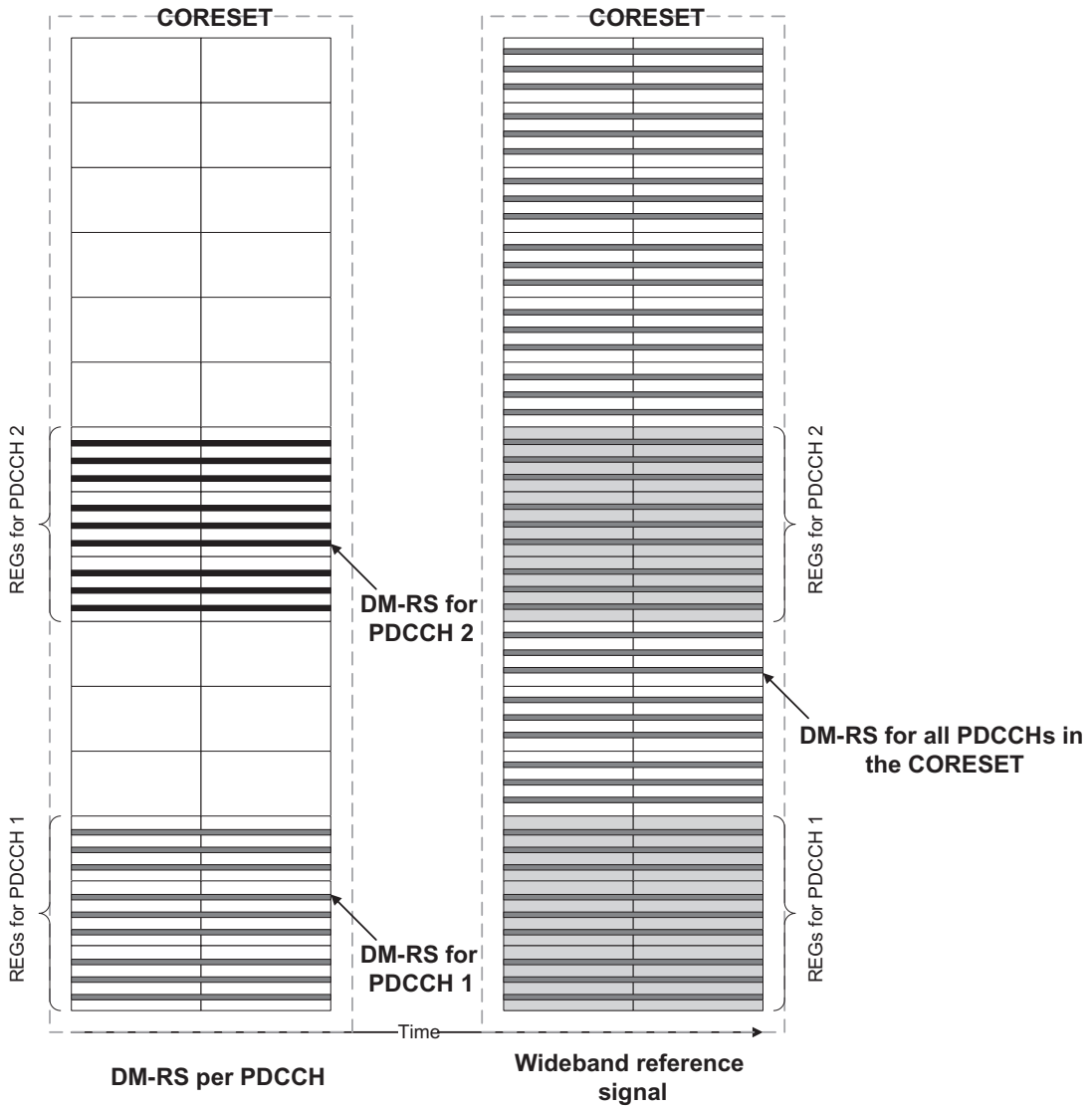


Figure 4.19
Illustration of the regular and wideband reference signals for PDCCH [68].

$d(0), d(1), \ldots, d(N_{symb} - 1)$. The block of complex-valued symbols $d(0), d(1), \ldots, d(N_{symb} - 1)$ is scaled by a factor of β_{PDCCH} and is mapped to resource elements (k, l), which are designated to PDCCH to be monitored by the UE and are not used for the associated PDCCH DM-RS, in increasing order of first k (frequency index) and l (time index) [6]. Fig. 4.20 illustrates the CORESET structures in overlapped and non-overlapped BWPs.

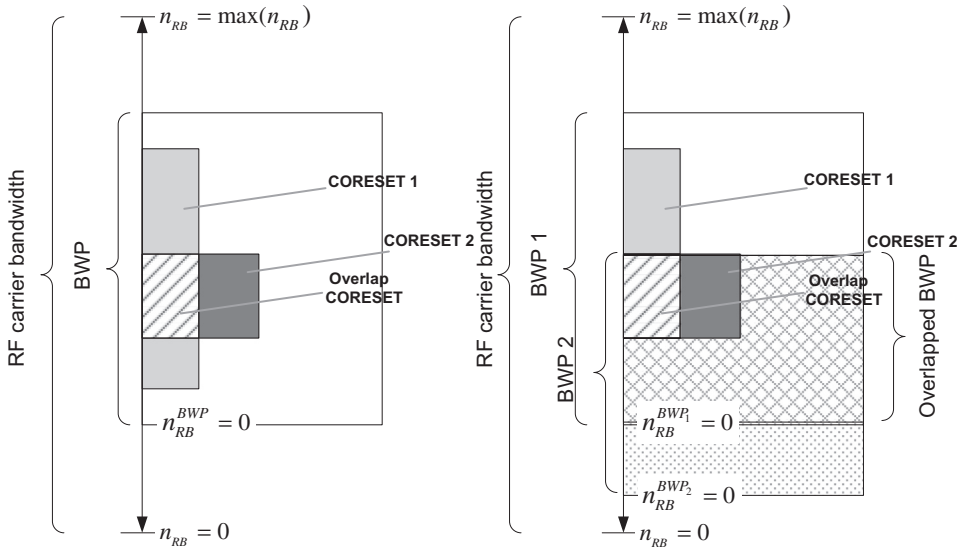


Figure 4.20
Illustration of the CORESET structures in overlapped and non-overlapped BWPs [71].

The PDCCH DM-RS sequence $r_{PDCCH}(l,m)$ on OFDM symbol l is defined by $r_{PDCCH}(l,m) = ([1-2c(2m)] + j[1-2c(2m+1)])/\sqrt{2}$ where $c(i)$ denotes a length-31 Gold sequence generated by the pseudo-random sequence generator defined in [6] and initialized with $c_{init} = \left[2^{17}\left(N_{slot}^{symbol} n_{slot} + l + 1\right)(2N_{ID}+1) + 2N_{ID}\right] \bmod 2^{31}$ where n_{slot} is the slot number within a frame, and $N_{ID} = \{0, 1, \ldots, 65535\}$ is signaled via the RRC parameter *pdcch-DMRS-ScramblingID*; otherwise $N_{ID} = N_{ID}^{cell}$. The PDCCH DM-RS sequence $r_{PDCCH}(l,m)$ is mapped to resource elements (k,l) such that $a(k,l) = \beta_{DMRS}^{PDCCH} r_{PDCCH}(3n+k', l); \forall n \in \mathbb{N}; k = nN_{sc}^{RB} + 4k' + 1;$ and $k' = 0, 1, 2$ provided that the following two conditions are met: (1) resource elements are within the REGs constituting the PDCCH that the UE attempts to decode, if the RRC parameter *precoderGranularity* is equal *sameAsREG-bundle*; and (2) all REGs are within the set of contiguous resource blocks in the CORESET where the UE attempts to decode PDCCH, if the RRC parameter *precoderGranularity* is equal to the size of the CORESET in the frequency domain [6]. The reference point for k (frequency index) is subcarrier 0 of the lowest numbered common resource block in the CORESET, if the CORESET is configured by the PBCH or by the *controlResourceSetZero* field in the *PDCCH-ConfigCommon* information element; otherwise, subcarrier 0 in common resource block 0. The parameter l is the OFDM symbol number within the slot. In the absence of CSI-RS configuration, the PDCCH DM-RS and SS/PBCH blocks are quasi-co-located with respect to Doppler shift, Doppler spread, average delay, delay spread, and spatial RX parameters.

The notion of wideband DM-RS has been introduced to assist the NR UE in channel estimation for the control channel detection (see Fig. 4.19). For each CORESET, the precoder granularity in the frequency domain is configurable between REG bundle size and the number of contiguous RBs in the frequency domain within the CORESET. The UE may assume DM-RS is present in all REGs within the set of contiguous RBs of the CORESET where and when at least one REG of a candidate is mapped [6].

4.1.3.2.2 UE Group-Common Signaling

An NR gNB can simultaneously support diverse service categories and thus mitigation of the performance degradation of interrupted services is an important issue in the physical layer design. While the flexible frame structure may alleviate this problem, due to the implementation complexity and unpredictability of URLLC packet arrival times, a more elaborate solution is needed in real-world deployment scenarios. 3GPP has considered a number of solutions in the course of NR development. For infrequent URLLC transmissions, one can give priority to URLLC transmissions while ensuring the reliability of the other transmissions interrupted by URLLC traffic. A preemption indicator transmitted by the base station indicates which resources are used for the URLLC transmission. If the URLLC packet is stretched in the frequency domain, the URLLC transmission will interrupt the entire system bandwidth and thus degrade all data channels in use. To notify scheduled users of this event, the base station broadcasts a preemption indicator consisting of time and/or frequency information of the interrupted resources. This indicator helps users identify the reason for packet errors and what part of the packet is unaffected from the interruption. Retransmission of selected code blocks when the ongoing service is interrupted by the URLLC transmission is another solution, where part of the code block that has been affected by URLLC transmission is retransmitted. By transmitting a combining indicator or flush-out indicator, the receiver can perform the soft symbol combining of the transmitted and retransmitted code blocks. One can further achieve better coding gain by lowering the code rate of the retransmitted code block.

An efficient scheduling scheme in terms of resource allocation and latency is to multiplex data with different transmission time lengths (i.e., mini-slots and slots) and in case of resource limitations to let the high-priority service use resources from lower priority service. This type of multiplexing is also referred to as preemption. For example, in NR downlink, a mini-slot carrying high-priority or delay-sensitive data can preempt an already ongoing slot-based transmission on the first available OFDM symbols without waiting for the next free resource. This operation enables ultra-low latency for mini-slot-based transmission, especially in the scenario where a long slot-based transmission has already been scheduled. A similar concept is also considered for the uplink and in general for LTE. At the cost of degrading the longer transmission, no additional resources need to be reserved in advance for the URLLC service. The impacted longer transmission is then

promptly repaired with a transmission containing a subset of the code block groups in a later transmission time, after providing the essential information to clean the corrupted soft values in the receive buffer from the preempted data. If the URLLC transmission occurs frequently, the efficiency of the above approach will be reduced due to the frequent retransmissions. To ensure the reliability of the ongoing services while supporting the URLLC transmission, robustness improvement and service sharing strategies may be adopted [68].

An NR UE can be configured to monitor group-common signaling via DCI format 2_1 which carries preemption indication related to multiplexing eMBB and URLLC traffic with different transmission durations in the downlink. Upon reception of preemption indication, a UE should apply an appropriate HARQ combining mechanism to retrieve the data despite the missing portion due to preemption. Group-common PDCCH is designed for the purpose of signaling to a group of UEs. It carries dynamic SFI, that is, via DCI format 2_0, to indicate which symbols in a slot are designated as downlink, uplink, or flexible symbols. The SFI carries an index to a UE-specific table containing permissible slot configurations [6]. The downlink preemption indicator is signaled via DCI format 2_1; however, whether a UE needs to monitor preemption indication is configured through RRC signaling. The UE is additionally configured with a set of serving cells; a mapping for each serving cell in the set of serving cells to the corresponding fields in DCI format 2_1; an information payload size for DCI format 2_1; and a bitmap for identification of punctured time-frequency resources via higher layer signaling. If the UE detects a DCI format 2_1 for a serving cell from the configured set of serving cells, the UE may assume that there is no transmission assigned to the UE in PRBs and symbols within the active downlink BWP, from a set of PRBs and a set of symbols of the last monitoring period, that are indicated by DCI format 2_1. Note that DCI format 2_1 indication is not applicable to the reception of SS/PBCH blocks [8].

A UE needs to monitor preemption indication carried by DCI format 2_1, if it is provided with RRC parameter *DownlinkPreemption* and it is configured with an INT-RNTI provided by RRC parameter *int-RNTI*. In that case, if the UE detects DCI format 2_1, the set of symbols indicated by a field in DCI format 2_1 includes the last $N_{slot}^{symb} T_{INT} 2^{\mu - \mu_{INT}}$ symbols prior to the first symbol of the CORESET in the slot. The parameter T_{INT} is the PDCCH monitoring periodicity provided by a higher layer parameter, N_{slot}^{symb} is the number of symbols per slot, μ is the subcarrier spacing configuration for a serving cell with mapping to a respective field in the DCI format 2_1, and μ_{INT} is the subcarrier spacing configuration of the downlink BWP where the UE receives the PDCCH conveying the DCI format 2_1. If the UE is configured with RRC parameter *TDD-UL-DL-ConfigurationCommon*, the symbols designated as uplink by the latter parameter are excluded from the last $N_{slot}^{symb} T_{INT} 2^{\mu - \mu_{INT}}$ symbols prior to the first symbol of the CORESET in the slot. The resulting set of symbols includes a number of symbols that is denoted as N_{INT} [8].

The UE is further provided with the (preemption) indication granularity for the set of PRBs and OFDM symbols (within a slot) that might be preempted through RRC parameter *timeFrequencySet*. If the value of latter parameter is zero (time-domain preemption region boundary), the 14-bit bitmap in DCI format 2_1 has a one-to-one correspondence with 14 groups of consecutive symbols from the set of symbols where each of the first $N_{INT} - 14\lfloor N_{INT}/14 \rfloor$ symbol groups includes $\lceil N_{INT}/14 \rceil$ symbols, each of the last $14 - N_{INT} + 14\lfloor N_{INT}/14 \rfloor$ symbol groups includes $\lfloor N_{INT}/14 \rfloor$ symbols, where in the bitmap, a bit value of "0" indicates transmission to the UE in the corresponding symbol group and a bit value of "1" indicates no transmission to the UE in the corresponding symbol group [8,13]. Interpretation of the bitmap is configurable such that each bit represents one OFDM symbol in the time domain and the full bandwidth part, or two OFDM symbols in the time domain and one half of the bandwidth part. Furthermore, the monitoring periodicity of the preemption indicator is configured in the device every *n*th slot. An example is shown in Fig. 4.21 where UE1 has been scheduled with a downlink transmission spanning one slot. During the transmission to UE1, delay-sensitive data for UE2 arrives at the gNB, which immediately schedules a transmission to UE2. Typically, if there are frequency resources available, the transmission to UE2 is scheduled using resources not overlapping with the ongoing transmission to UE1. However, in a heavy-loaded network, this may not be possible and there is no option but to use some of the resources originally allocated to UE1 for the delay-sensitive transmission to UE2. We refer to this case as the transmission to UE2 preempting the transmission to UE1, which would experience temporary performance degradation because some of the resources that UE1 assumes to contain its own data contain data for UE2.

If the value of the RRC parameter *timeFrequencySet* is one (frequency-domain preemption region boundary) then seven pairs of bits of a field in the DCI format 2_1 have a one-to-one mapping with seven groups of consecutive symbols where each of the first $N_{INT} - 7\lfloor N_{INT}/7 \rfloor$ symbol groups includes $\lfloor N_{INT}/7 \rfloor$ symbols, and each of the last $7 - N_{INT} + 7\lfloor N_{INT}/7 \rfloor$ symbol groups includes $\lfloor N_{INT}/7 \rfloor$ symbols (as shown in Fig. 4.21). The first bit in a pair of bits for a symbol group is applicable to the subset of first $\lfloor B_{INT}/2 \rfloor$ PRBs from the set of B_{INT} PRBs, and the second bit in the pair of bits for the symbol group is applicable to the subset of last $\lfloor B_{INT}/2 \rfloor$ PRBs from the set of B_{INT} PRBs, a bit value of "0" indicates transmission to the UE in the corresponding symbol group and subset of PRBs, and a bit value of "1" indicates no transmission to the UE in the corresponding symbol group and subset of PRBs [8,13].

In relation to the dynamic slot configuration, if a UE is configured by RRC parameter *SlotFormatIndicator*, it will be provided with SFI-RNTI via higher layer parameter *sfi-RNTI,* and the payload size of DCI format 2_0 via RRC parameter *dci-PayloadSize*. The UE is also provided, in one or more serving cells, with a configuration for search space set S and the corresponding CORESET P for monitoring the first $M_{P,S}^{L_{SFI}}$ PDCCH candidates

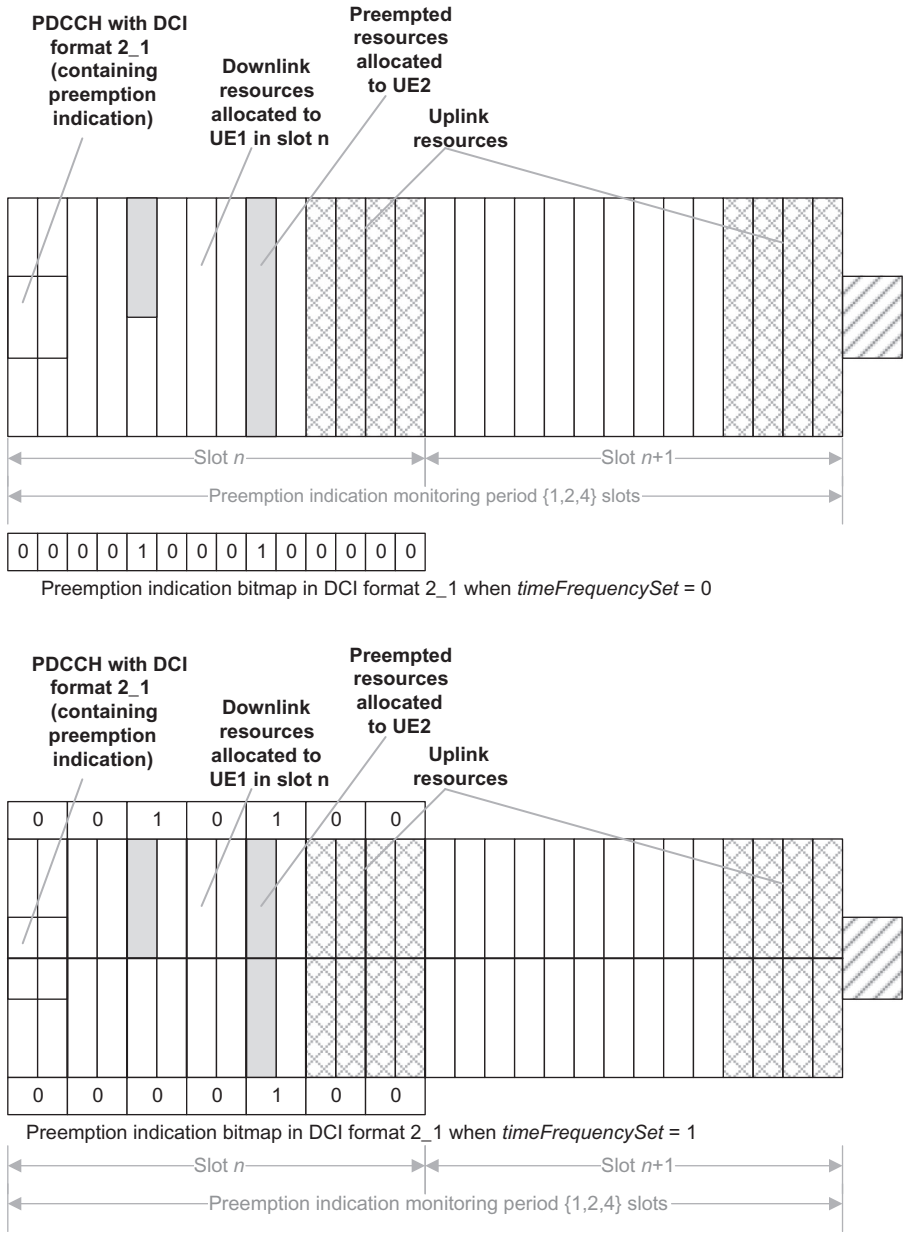


Figure 4.21
Illustration of downlink preemption indication (interrupted transmission indication) [8].

related to DCI format 2_0 at aggregation level of L_{SFI} CCEs to locate Type3-PDCCH common search space. The DCI format 2_0 indicates a slot format corresponding to each slot in the downlink or uplink BWPs of a serving cell. The indication is communicated by setting the value of the SFI index in DCI format 2_0 for the serving cell to a combination of slot

formats for a number of slots. A slot format is identified by its corresponding index and the mapping between values of the SFI index and combinations of slot formats is signaled via RRC parameters [8].

In TDD operation mode, the UE is provided, via RRC signaling, with a reference subcarrier spacing configuration μ_{SFI} for each slot format in a set of slot formats indicated by the SFI index in DCI format 2_0. An active downlink and uplink BWP pair are associated with subcarrier spacing configuration $\mu \geq \mu_{SFI}$. Each slot format in the combination of slot formats is applicable to $2^{(\mu-\mu_{SFI})}$ consecutive slots in the active downlink and uplink BWP pair where the first slot starts at the same time as the first slot for the reference subcarrier spacing configuration and each downlink, uplink, or flexible symbol for the reference subcarrier spacing configuration corresponds to $2^{(\mu-\mu_{SFI})}$ consecutive downlink, uplink, or flexible symbols for the subcarrier spacing configuration μ [8]. In FDD operation mode, the SFI index field in DCI format 2_0 indicates an assortment of slot formats that include (separate) combinations of slot formats for reference downlink and uplink BWPs of the serving cell. The UE is provided via RRC signaling with the reference subcarrier spacing configurations μ_{SFI-DL} or μ_{SFI-UL} for the combination of slot formats indicated by the SFI index field value in DCI format 2_0 for the reference downlink and uplink BWPs of the serving cell, respectively.

4.1.3.2.3 Downlink Control Information Formats

The PDCCH payload is known as DCI to which a 24-bit CRC is attached to detect transmission errors and to aid the decoder in the UE receiver. Compared to LTE, the CRC size has been increased to reduce the risk of incorrectly received control information and to assist early termination of the decoding operation in the receiver. A DCI transports downlink and uplink scheduling information, requests for aperiodic channel quality indicator (CQI) reports, or uplink power control commands for one cell and one RNTI. Depending on the content and purpose of each DCI, different formats are defined. Each DCI payload is processed by information element multiplexing, CRC attachment, channel coding, and rate matching. The DCI formats defined in NR are shown in Table 4.4 along with their use cases. Similar to LTE, the UE identity modifies the CRC transmitted through a scrambling operation. When the DCI is received, the UE will calculate a scrambled CRC on the payload part using the same transmit-side procedure and compares it against the received CRC. If the CRC matches, the message is declared to be correctly received and intended for the UE. Therefore, the identity of the UE that is supposed to receive the DCI message is implicitly encoded in the CRC, which reduces the number of bits necessary to transmit on the PDCCH. Note that the RNTI which is further scrambled with the DCI CRC is not necessarily the identity of the device (in the case of C-RNTI), rather it can be different type of group or common RNTIs used to indicate paging or a random-access response.

The information fields in the DCI formats shown in Table 4.4 are mapped to the information bits a_0 to $a_{N_{DCI}-1}$ such that the first field is mapped to the lowest order information bit

Table 4.4: NR downlink control information (DCI) formats [7].

DCI Format	Purpose	Application
0_0	Uplink Scheduling	Scheduling of PUSCH in one cell
0_1		Scheduling of PUSCH in one cell
1_0	Downlink scheduling	Scheduling of PDSCH in one cell
		DCI format 1_0 with CRC scrambled by C-RNTI
		DCI format 1_0 with CRC scrambled by RA-RNTI
		DCI format 1_0 with CRC scrambled by TC-RNTI
		DCI format 1_0 with CRC scrambled by SI-RNTI
		DCI format 1_0 with CRC scrambled by P-RNTI
1_1		Scheduling of PDSCH in one cell
2_0	Other purposes	Notifying a group of UEs of the slot format
2_1		Notifying a group of UEs of the PRB(s) and OFDM symbol(s) where UE may assume that no transmission is intended for the UE
2_2		Transmission of TPC commands for PUCCH and PUSCH
2_3		Transmission of a group of TPC commands for SRS transmissions by one or more UEs

a_0, and each successive field is mapped to higher order information bits. If the number of information bits in a DCI format is less than 12 bits, zero padding is used until the payload size is equal to 12. The DCI size of the uplink DCI format 0_1 and downlink DCI format 1_1 are made equal with padding bits added to the smaller of the two in order to reduce the number of blind decoding. It may appear that parts of the DCI content are the same for the different formats; however, there are differences due to different capabilities supported by each DCI format. The content of various DCI formats is described in the following [7]. Note that the information fields and their interpretation may change according to the RNTI value that is used in conjunction with the DCI value.

DCI Format 0_0

- Identifier for DCI formats (1 bit): It is a bit to indicate whether the DCI is a downlink assignment or an uplink grant. The value of this bit is always set to 0, indicating an uplink DCI format.
- Frequency domain resource assignment: The number of bits for this field is determined by following formula $\lceil \log_2(N_{RB}^{UL,BWP}(N_{RB}^{UL,BWP}+1)/2) \rceil$. The meaning of $N_{RB}^{UL,BWP}$ varies depending on the search space where DCI format 0_0 is expected to be detected. When transmitted in common search space, it indicates the size of the initial bandwidth part, whereas in UE-specific search space, it would indicate the size of the active bandwidth part, if the following criteria are satisfied: the total number of different DCI sizes monitored per slot is less than 4 and the total number of different DCI sizes with C-RNTI

monitored per slot is less than 3. The value of this field is determined in two cases: If PUSCH hopping is enabled and the resource allocation is Type 1, N_{UL-hop} MSB bits are used to indicate the frequency offset. If PUSCH hopping is disabled and the resource allocation is type 1, the entire bits of this field would indicate PUSCH RIV.

- Time-domain resource assignment (4 bits): It carries the row index of the items in *pusch-TimedomainAllocationList* in RRC message for PUSCH configuration, where the indexed row defines the slot offset K_2, the start and length indicator *SLIV*, and the PUSCH mapping type to be applied in the PUSCH transmission.
- Frequency hopping flag (1 bit): It is used to handle frequency hopping for resource allocation Type 1.
- Modulation and coding scheme (5 bits): It is used to provide the device with information about the modulation scheme, the code rate, and the transport block size (TBS) (see Table 4.13).
- New data indicator (1 bit): It is used to indicate whether the grant relates to retransmission of a TB or transmission of a new TB.
- Redundancy version (RV) (2 bits): This field determines the value of the parameter $rv_{id} = 0, 1, 2, 3$ which is used to indicate the redundant information sent in a HARQ retransmission.
- HARQ process number (4 bits): It informs the device about the HARQ process to be used for soft combining.
- Transmit power control (TPC) command for scheduled PUSCH (2 bits): It is used to adjust the PUSCH transmission power.
- Uplink/Supplementary uplink (SUL) indicator (0 or 1 bit): One bit to indicate whether the grant relates to the SUL or the ordinary uplink, for UEs configured with SUL in the cell. It is only present if a SUL is configured as part of the SI.

DCI Format 0_1

- Identifier for DCI formats (1 bit): A bit to indicate whether the DCI is a downlink assignment or an uplink grant. The value of this bit is always set to 0, indicating an uplink DCI format.
- Carrier indicator (0 or 3 bits): This field is present if cross-carrier scheduling is configured and is used to indicate to which component carrier the DCI is related.
- Bandwidth part indicator (0, 1, 2 bits): It is used to activate one of up to four bandwidth parts configured by higher layer signaling. It is determined by the number of uplink BWPs n_{BWP} configured via RRC signaling, excluding the initial uplink bandwidth part. The size of this field is $\lceil \log_2(n_{BWP}) \rceil$ bits.
- Frequency domain resource assignment: This field indicates the resource blocks on one component carrier over which the device should transmit PUSCH. The number of bits is variable and dependent on the resource allocation type. The number of bits is equal to N_{RBG} bits, if resource allocation Type 0 is configured for the UE. For resource

allocation Type 1, the number of bits is equal to $\lceil \log_2(N_{RB}^{UL,BWP}(N_{RB}^{UL,BWP}+1)/2) \rceil$ bits or $\max\left(\lceil \log_2(N_{RB}^{UL,BWP}(N_{RB}^{UL,BWP}+1)/2) \rceil, N_{RBG}\right)+1$ bits if both resource allocation Type 0 and 1 are configured. Note that if both resource allocation Type 0 and 1 are configured, the MSB bit is used to distinguish resource allocation Type 0 from Type 1.

- Time domain resource assignment (0, 1, 2, 3, or 4 bits): This field indicates the resource allocation in the time domain. The number of bits is determined as $\lceil \log_2(I) \rceil$ bits, where I denotes the number of entries in RRC parameter *pusch-TimeDomainAllocationList.*
- VRB-to-PRB mapping (1 bit): It is used to indicate whether interleaved or non-interleaved VRB-to-PRB mapping should be used.
- Frequency hopping flag (1 bit): It is used to handle frequency hopping for resource allocation Type 1.
- Modulation and coding scheme (5 bits): It is used to provide the device with information about the modulation scheme, the code rate, and the TBS (see Table 4.13).
- New data indicator (1 bit): It is used to indicate whether the grant relates to retransmission of a TB or transmission of a new TB.
- Redundancy version (2 bits): This field determines the value of the parameter $rv_{id} = 0, 1, 2, 3$ which is used to indicate the redundant information sent in a HARQ retransmission.
- HARQ process number (4 bits): It informs the device about the HARQ process to be used for soft combining.
- First downlink assignment index (DAI) (1 or 2 bits): The DAI is used for handling of HARQ codebooks when UCI is transmitted on PUSCH. This field would be one bit for semi-static HARQ-ACK codebook and 2 bits for dynamic HARQ-ACK codebook.
- Second DAI (0 or 2 bits): This field would be 2 bits for dynamic HARQ-ACK codebook with two HARQ-ACK subcodebooks and zero bit otherwise.
- TPC command for scheduled PUSCH (2 bits): It is used to adjust the PUSCH transmission power.
- Sounding reference signals (SRS) resource indicator: The SRI is used to determine the antenna ports and uplink transmission beam to use for PUSCH transmission. The number of bits depends on the number of SRS groups configured and whether codebook-based or non-codebook-based precoding is used.
- Precoding information and number of layers (0, 2, 3, 4, 5, or 6 bits): This field is used to select the precoding matrix **W** and the number of layers for codebook-based precoding. The number of bits depends on the number of antenna ports and the maximum rank supported by the UE.
- Antenna ports (2, 3, 4, or 5 bits): This field indicates the antenna ports on which the data are transmitted as well as antenna ports that are scheduled for other users.
- SRS request (2 bits): This field is used to request transmission of a sounding RS.

- CSI request (0, 1, 2, 3, 4, 5, or 6 bits): This field is used to request transmission of a CSI report.
- Code block group (CBG) transmission information (0, 2, 4, 6, or 8 bits): This field indicates the code block groups for retransmission.
- PT-RS-DM-RS association (0 or 2 bits): This field is used to indicate the association between the DM-RS and PT-RS ports.
- *beta_offset* Indicator (0 or 2 bits): This information is used to control the amount of resources used by UCI on PUSCH in case dynamic beta offset signaling is configured for DCI format 0_1.
- DM-RS sequence initialization (0 or 1 bit): This information is used to select between two preconfigured initialization values for the DM-RS sequence. It would be zero bit, if the transform precoder is enabled and 1 bit, otherwise.
- Uplink/SUL indicator (0 or 1 bit): It would be zero bit for UEs that are not configured with SUL in the cell or UEs configured with SUL in the cell but only physical uplink control channel (PUCCH) carrier in the cell is configured for PUSCH transmission; otherwise, 0 bit for UEs configured with SUL in the cell.

DCI Format 1_0

- Identifier for DCI formats (1 bit): A bit to indicate whether the DCI is a downlink assignment or an uplink grant. The value of this field is always set to 1, indicating a downlink DCI format.
- Frequency-domain resource assignment: This field indicates the resource blocks on one component carrier on which the UE receives PDSCH. The size of this field depends on the size of the bandwidth part and on the resource allocation type, that is, Type 0 only, Type 1 only, or dynamic switching between the two types. The number of bits is $\lceil \log_2(N_{RB}^{DL,BWP}(N_{RB}^{DL,BWP}+1)/2) \rceil$ where the interpretation of $N_{RB}^{DL,BWP}$ depends on the search space where DCI format 1_0 is monitored. The parameter $N_{RB}^{DL,BWP}$ indicates the size of the active downlink bandwidth part, if DCI format 1_0 is monitored in the UE-specific search space and if the total number of different DCI sizes configured to monitor is less than 4 and the total number of different DCI sizes with C-RNTI configured to monitor is less than 3 for the cell; otherwise, $N_{RB}^{DL,BWP}$ would denote the size of CORESET 0.
- Time-domain resource assignment (4 bits): This field indicates the resource allocation in the time domain. When the UE is scheduled to receive PDSCH by a DCI, the time-domain resource assignment field of the DCI provides a row index of a higher layer-configured table *pdsch-symbolAllocation*, where the indexed row defines the slot offset K_0, the start and length indicator *SLIV*, and the PDSCH mapping type for PDSCH reception.
- VRB-to-PRB mapping (1 bit): It indicates whether interleaved or non-interleaved VRB-to-PRB mapping should be used and only presents for resource allocation Type 1.

- Modulation and coding scheme (5 bits): It is used to provide the UE with information about the modulation scheme, code rate, and TBS (see Table 4.13).
- New data indicator (1 bit): It is used to clear the UE soft buffer for initial transmissions.
- Redundancy version (2 bits): This field determines the value of the parameter $rv_{id} = 0,\ 1,\ 2,\ 3$ which is used to indicate the redundant information sent in a HARQ retransmission.
- HARQ process number (4 bits): This informs the device about the HARQ process to use for soft combining.
- Downlink assignment index (2 bits): The DAI only presents when a dynamic HARQ codebook is used. The DCI format 1_1 supports 0, 2, or 4 bits, while DCI format 1_0 uses 2 bits.
- TPC command for scheduled PUCCH (2 bits): It is used to adjust the PUCCH transmission power.
- PUCCH resource indicator (3 bits): It is used to select PUCCH resources from a set of configured resources.
- PDSCH-to-HARQ feedback timing indicator (3 bits): It provides information on when the HARQ acknowledgment should be transmitted relative to the PDSCH transmission.

DCI Format 1_1

- Identifier for DCI formats (1 bit): The value of this bit is always set to 1, indicating a downlink DCI format.
- Carrier indicator (0 or 3 bits): This field is present if cross-carrier scheduling is configured and is used to indicate the component carrier that the DCI corresponds to.
- Bandwidth part indicator (0, 1, or 2 bits): The number of bits is determined by the number of downlink BWPs $n_{BWP,RRC}$ configured by higher layers, excluding the initial downlink bandwidth part. The size of this field is equal to $\lceil \log_2(n_{BWP}) \rceil$ bits, where $n_{BWP} = n_{BWP,RRC} + 1$, if $n_{BWP,RRC} \leq 3$ in which case the bandwidth part indicator is equivalent to the higher layer parameter *BWP-Id*; otherwise $n_{BWP} = n_{BWP,RRC}$.
- Frequency-domain resource assignment: This field indicates the resource blocks on one component carrier on which the device should receive PDSCH. The number of bits is variable and dependent on the resource allocation type. The number of bits is equal to N_{RBG} bits, if resource allocation Type 0 is configured for the UE. For resource allocation Type 1, the number of bits is equal to $\lceil \log_2(N_{RB}^{UL,BWP}(N_{RB}^{UL,BWP}+1)/2) \rceil$ bits or $\max\left(\lceil \log_2(N_{RB}^{DL,BWP}(N_{RB}^{DL,BWP}+1)/2) \rceil, N_{RBG}\right) + 1$ bits, if both resource allocation Types 0 and 1 are configured. Note that if both resource allocation Types 0 and 1 are configured, the MSB bit is used to distinguish resource allocation Type 0 from Type 1.
- Time-domain resource assignment (0, 1, 2, 3, or 4 bits): The size of this field is determined as $\lceil \log_2(I) \rceil$ bits, where I is the number of entries in the higher layer parameter *pdsch-TimeDomainAllocationList*. When the UE is scheduled to receive PDSCH by a

DCI, the time-domain resource assignment field of the DCI provides a row index of a higher layer-configured table *pdsch-symbolAllocation*, where the indexed row defines the slot offset K_0, the start and length indicator *SLIV*, and the PDSCH mapping type for PDSCH reception.

- VRB-to-PRB mapping (0 or 1 bit): It indicates whether interleaved or non-interleaved VRB-to-PRB mapping should be used and only presents for resource allocation Type 1.
- PRB bundling size indicator (0 or 1 bit): It is used to indicate the PDSCH bundling size. It is zero bit if the RRC parameter *prb-BundlingType* is not configured or is set to "static." It is one bit, if the higher layer parameter *prb-BundlingType* is set to "dynamic."
- Rate matching indicator (0, 1, or 2 bits): The number of bits is determined according to RRC parameters *rateMatchPatternGroup1* and *rateMatchPatternGroup2*.
- Modulation and coding scheme (TB 1) (5 bits): It is used to provide the UE with information about the modulation scheme, code rate, and TBS (see Table 4.13) related to the first transport block.
- New data indicator (TB 1) (1 bit): It is used to clear the UE soft buffer for initial transmissions related to the first transport block.
- Redundancy version (TB 1) (2 bits): This field determines the value of the parameter which is used to indicate the redundant information sent in a HARQ retransmission related to the first transport block.
- Modulation and coding scheme (TB 2)[5] (5 bits): It is used to provide the UE with information about the modulation scheme, code rate, and TBS related to the second transport block (see Table 4.13).
- New data indicator (TB 2) (1 bit): It is used to clear the UE soft buffer for initial transmissions related to the second transport block.
- Redundancy version (TB 2) (2 bits): This field determines the value of the parameter which is used to indicate the redundant information sent in a HARQ retransmission related to the second transport block.
- HARQ process number (4 bits): This informs the device about the HARQ process to use for soft combining.
- Downlink assignment index (0, 2, or 4 bits): The number of bits is 4, if more than one serving cell is configured in the downlink and the RRC parameter *pdsch-HARQ-ACK-Codebook = dynamic*, where the two MSB bits are the counter DAI and the two LSB bits are the total DAI. The number of bits is 2, if only one serving cell is configured in the downlink and the RRC parameter *pdsch-HARQ-ACK-Codebook = dynamic*, where the 2 bits are the counter DAI; zero bits otherwise.

[5] If a second transport block is present (only if more than four layers of spatial multiplexing are supported in DCI format 1_1), the three fields above are repeated for the second transport block.

- TPC command for scheduled PUCCH (2 bits): It is used to adjust the PUCCH transmission power.
- PUCCH resource indicator (2 bits): It is used to select PUCCH resources from a set of configured resources.
- PDSCH-to-HARQ_Feedback timing indicator (3 bits): It provides information on when the HARQ acknowledgment should be transmitted relative to the PDSCH transmission.
- Antenna port(s) and number of layers (4, 5, or 6 bits): The antenna ports $\{p_0, \ldots, p_{N_A-1}\}$ are determined according to the ordering of DM-RS port(s). If a UE is configured with both *dmrs-DownlinkForPDSCH-MappingTypeA* and *dmrs-DownlinkForPDSCH-MappingTypeB*, the size of this field is equal to $\max(x_A, x_B)$, where x_A and x_B are the "antenna ports" bit sizes derived from *dmrs-DownlinkForPDSCH-MappingTypeA* and *dmrs-DownlinkForPDSCH-MappingTypeB*, respectively. A number of zeros are inserted in the $|x_A - x_B|$ MSB positions of this field, if the mapping type of the PDSCH corresponds to the smaller value of x_A or x_B.
- Transmission configuration indication (0 or 3 bits): The size of this field is zero bit, if RRC parameter *tci-PresentInDCI* is not enabled; otherwise it would carry 3 bits.
- SRS request (2 or 3 bits): It is used to request transmission of a sounding reference signals in the uplink. For UEs not configured with SUL in the cell, 2 bits are used, whereas for UEs that are configured with SUL in the cell, 3 bits are used where the first bit is the non-SUL/SUL indicator and the second and third bits are used to request periodic or aperiodic SRS transmission. This bit field may also indicate the associated CSI-RS.
- CBG transmission information (0, 2, 4, 6, or 8 bits): If CBG retransmissions are configured, this field indicates the code block groups that are retransmitted.
- CBG flushing out information (0 or 1): If CBG retransmissions are configured, the content of this field indicates the soft buffer flushing, which is determined by RRC parameter *codeBlockGroupFlushIndicator*.
- DM-RS sequence initialization (1 bit): This information is used to select between two preconfigured initialization values for the DM-RS sequence. It would be zero bit, if the transform precoder is enabled and 1 bit otherwise.

DCI Format 2_0 DCI format 2_0 is used for notifying the UE of slot format. The SFI is transmitted using regular PDCCH structure and SFI-RNTI, which is common to a group of UEs. To assist the device in the blind decoding, the device is configured with information about the up to two PDCCH candidates on which the SFI can be transmitted. DCI format 2_0 with CRC scrambled with SFI-RNTI carries *Slot format indicator* 1, *Slot format indicator* $2, \ldots,$ *Slot format indicator* N. The size of DCI format 2_0 is configurable by higher layers up to 128 bits.

DCI Format 2_1 DCI format 2_1 is used to signal the preemption indication to the device. It is transmitted using the regular PDCCH structure, using INT-RNTI which can be common to multiple devices. In other words, DCI format 2_1 is used for notifying the UEs of

the PRB(s) and OFDM symbol(s) that are preempted and have no transmission intended for the UE. DCI format 2_1 with CRC scrambled by INT-RNTI carries *Preemption indication* 1, *Preemption indication* 2, . . ., *Preemption indication N*. The size of DCI format 2_1 is configurable by higher layers up to 126 bits where each preemption indication is 14 bits.

DCI Format 2_2 The main purpose of DCI format 2_2 is to support power control for semi-persistent scheduling (SPS) since there is no dynamic scheduling assignment or scheduling grant which can include the power control information for PUCCH and PUSCH in this case. The power-control message is addressed to a group of UEs using an RNTI specific for that group and each UE is configured with the power control bits in the message. DCI format 2_2 is further aligned with the size of DCI formats 0_0/1_0 to reduce the blind decoding complexity. DCI format 2_2 with CRC scrambled by TPC-PUSCH-RNTI or TPC-PUCCH-RNTI carries *block number* 1, *block number* 2, . . .*block number N*. The RRC parameters *tpc-PUSCH* or *tpc-PUCCH* determine the index to the block number for a cell uplink, with the following fields defined for each block: (1) closed-loop indicator (0 or 1 bit) and (2) TPC command (2 bits).

DCI Format 2_3 DCI format 2_3 is used for power control of uplink sounding reference signals for the UEs which have not linked the SRS power control to the PUSCH power control, either because independent control was desirable, or the UE was configured without PUCCH and PUSCH. DCI format 2_3 structure is similar to DCI format 2_2, with the possibility to individually configure 2 bits for SRS request in addition to the two power control bits. DCI format 2_3 is aligned with the size of DCI formats 0_0/1_0 to reduce the blind decoding complexity. DCI format 2_3 with CRC scrambled by TPC-SRS-RNTI carries *block number* 1, *block number* 2, . . ., *block number N* where the starting position of a block is determined by the parameter *startingBitOfFormat2-3* provided by the higher layers for the UE configured with the block. If the UE is configured with RRC parameter *srs-TPC-PDCCH-Group = typeA* for an uplink carrier without PUCCH and PUSCH or when the SRS power control is not linked to PUSCH power control, in DCI format 2_3 one block is configured for the UE containing 0 or 2 bits of SRS request. The TPC commands *TPC command number* 1, *TPC command number* 2, . . ., *TPC command number N* apply to the respective carriers. If the UE is configured with RRC parameter *srs-TPC-PDCCH-Group = typeB* for an uplink carrier without PUCCH and PUSCH or an uplink carrier on which the SRS power control is not tied to PUSCH power control, one or more blocks are configured for the UE by the higher layers. In that case, each block applies to an uplink carrier and DCI format 2_3 contains 0 or 2 bits of SRS request and 2 bits of TPC command.

4.1.3.2.4 Common and UE-Specific Search Spaces

A UE may be configured with one or more CORESETs (using UE-specific or common signaling) with a maximum of three CORESETs per configured downlink BWP. Note that the

scheduling flexibility may not be impacted by limiting the maximum number of CORESETs since different monitoring occasions can be configured flexibly even in association with the same CORESET. It is important to further note that the concept of PDCCH monitoring periodicity is defined per search space set and is not configured at the CORESET-level. Every configured search space with a certain monitoring periodicity (in terms of slots and starting symbols within the monitored slots) is associated with a CORESET.

In LTE, the DCI format was closely coupled with the DCI size and monitoring for a certain DCI format in most cases implied monitoring for a new DCI size. In NR, the DCI formats and DCI sizes are decoupled. Different formats can have different DCI sizes, but several formats can share the same DCI size. This allows adding more formats in the future without increasing the number of blind decoding attempts. An NR device needs to monitor up to four different DCI sizes: one size used for the fallback DCI formats, one for downlink scheduling assignments, and unless the uplink downlink non-fallback formats are size-aligned, one for uplink scheduling grant. In addition, a device may need to monitor SFI and/or preemption indication DCIs using a fourth size, depending on the configuration.

An NR UE needs to monitor the PDCCH candidates at multiple aggregation levels for the detection and reception of PDCCH. Inside a configured CORESET, NR SS defines the PDCCH candidates of each AL [8]. In NR, PDCCH employs DM-RS-based transmission. Unlike LTE PDCCH where cell-specific reference signals were used for coherent demodulation, the NR channel estimation complexity scales with the number of CCEs being monitored. Thus it is important to balance the scheduling flexibility against UE implementation burden to facilitate cost-efficient UE implementation. For PDCCH DM-RS in a CORESET, the antenna port QCL configuration relating to the SS/PBCH block antenna port(s) or configured CSI-RS antenna port(s), is on a per-CORESET basis. This implies that in mmWave deployments, which rely on beam-sweeping operations, different CORESET and search space configurations corresponding to different received beams are necessary [8,53].

The CCE structure described in the previous section helps reduce the number of blind decoding attempts; however, it is required to have mechanisms to limit the number of PDCCH candidates that the device is expected to monitor. From a scheduling point of view, restrictions in the allowed aggregations are undesirable as they may reduce the scheduling flexibility and require additional processing at the transmitter side. At the same time, requiring the device to monitor all possible CCE aggregations in all configured CORESETs significantly increases device complexity and power consumption. A search space is a set of candidate control channels comprising a set of CCEs at a given aggregation level, which the device is supposed to monitor and decode. Due to multiple aggregation levels, a device can have multiple search

spaces. There can be multiple SSs using the same CORESET or multiple CORESETs configured for a device. A device is not expected to monitor PDCCH outside its active bandwidth part. At a configured monitoring occasion for a search space, the devices will attempt to decode the candidate PDCCHs for that search space. Five different aggregation levels corresponding to 1, 2, 4, 8, and 16 CCEs can be configured. The highest aggregation level is meant to support extreme coverage requirements [6,8,14].

The number of PDCCH candidates can be configured per search space and per aggregation level. When the UE attempts to decode a candidate PDCCH, the content of the control channel is declared as valid, if the CRC checks and the device can successfully process the contained information, that is, scheduling assignment, and uplink grants. If the CRC does not pass, the information is either subject to uncorrectable transmission errors or intended for another UE and in either case the device ignores that PDCCH transmission. The gNB can only address a UE, if the corresponding control information is transmitted on a PDCCH formed by the CCEs in one of the UE's search spaces. Therefore, for efficient utilization of the CCEs in the system, the UE should be associated with different search spaces. Each device in the system can be configured with one or more UE-specific search spaces. Since a UE-specific search space is typically smaller than the number of PDCCHs that the network can transmit at the corresponding aggregation level, there must be a mechanism to determine a set of CCEs in UE-specific search space. One option is to allow the network to configure the UE-specific search space for each device, in the same way that CORESETs are configured. However, this would require explicit signaling exchange with each device and possibly reconfiguration at handover. Instead, the UE-specific search spaces for PDCCH are defined without explicit signaling and based on the device unique identity in the cell in the connected mode, that is, C-RNTI. Furthermore, the set of CCEs that the device should monitor at a certain aggregation level varies as a function of time to avoid two devices constantly blocking each other. If they collide at one time instant, they are not likely to collide at the next time instant. In each of these search spaces, the UE attempts to decode the PDCCHs using the UE-specific C-RNTI. There is also information intended for a group of UEs in the cell. These messages are scheduled with different predefined RNTIs, for example, SI-RNTI for scheduling system information, P-RNTI for transmission of a paging message, RA-RNTI for transmission of the random-access response, TPC-RNTI for uplink power control, INT-RNTI for preemption indication, and SFI-RNTI for slot format configuration. As part of random-access procedure, it is necessary to transmit information to a device before it is assigned a unique identity. These types of information cannot rely on a UE-specific search space as different devices would monitor different CCEs despite the message being intended for all of them. Thus common search spaces are defined, where a common search space is similar in structure to a UE-specific search space with the difference that the set of CCEs is predefined and known to all devices irrespective of their own identity [14].

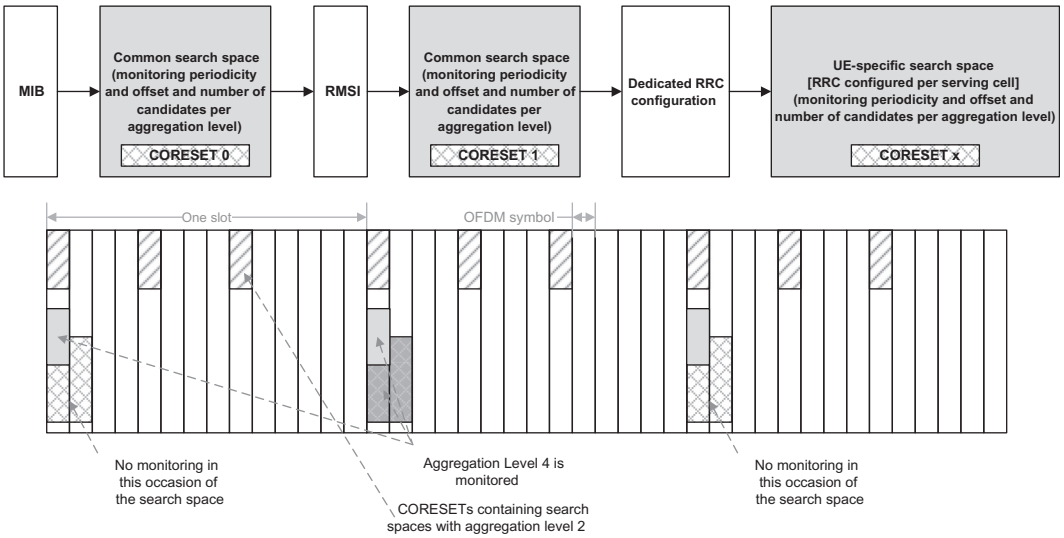


Figure 4.22
Procedure for PDCCH search space configuration and example search spaces [68].

The number of blind decoding attempts is proportional to subcarrier spacing and the slot duration. For 15/30/60/120 kHz subcarrier spacing, up to 44/36/22/20 blind decoding attempts per slot can be supported across all DCI payload sizes, respectively. It must be noted that the number of blind decoding attempts is not the only measure of UE complexity, channel estimation efforts also need to be taken in consideration. The number of channel estimations for subcarrier spacings of 15/30/60/120 kHz has been limited to 56/56/48/32 CCEs across all CORESETs in a slot [8]. Depending on the configuration, the number of PDCCH candidates may be limited either by the number of blind decoding attempts, or by the number of channel estimates. In order to minimize the device complexity, a device monitors a maximum of three different DCI sizes using C-RNTI and one DCI size using other RNTIs. In carrier aggregation scenarios, the general blind decoding operation described earlier is applied per component carrier. While the total number of channel estimates and blind decoding attempts increases compared to the single carrier case, there is no direct proportionality between the number of aggregated carriers and blind decoding attempts [14]. The procedure for PDCCH search space configuration and example search spaces are shown in Fig. 4.22.

As we stated earlier, a set of PDCCH candidates are defined for each UE to monitor, which are referred to as PDCCH search spaces. A search space can be categorized as common or UE-specific. In other words, a search space is defined by the PDCCH candidates that need to be monitored. These candidates are determined by a hashing function that operates within

a set of CCEs in a particular CORESET and the monitoring periodicity and offsets that determine when the search space should be monitored (see Fig. 4.22). The UE is required to monitor PDCCH candidates in one or more of the following search spaces [8]:

- *Type0-PDCCH* common search space set configured by *pdcch-ConfigSIB1* in *MasterInformationBlock* or by *searchSpaceSIB1* in *PDCCH-ConfigCommon* or by *searchSpaceZero* in *PDCCH-ConfigCommon* for a DCI format, the CRC of which is scrambled with SI-RNTI in the primary cell.
- *Type0A-PDCCH* common search space set configured by *searchSpaceOtherSystemInformation* in *PDCCH-ConfigCommon* for a DCI format, the CRC of which is scrambled with SI-RNTI in the primary cell.
- *Type1-PDCCH* common search space set configured by *ra-SearchSpace* in *PDCCH-ConfigCommon* for a DCI format, the CRC of which is scrambled with RA-RNTI, TC-RNTI, or C-RNTI in the primary cell.
- *Type2-PDCCH* common search space set configured by *pagingSearchSpace* in *PDCCH-ConfigCommon* for a DCI format, the CRC of which is scrambled with P-RNTI in the primary cell.
- *Type3-PDCCH* common search space set configured by *SearchSpace* in *PDCCH-Config* with *searchSpaceType = common* for a DCI format, the CRC of which is scrambled with INT-RNTI, SFI-RNTI, TPC-PUSCH-RNTI, TPC-PUCCH-RNTI, TPC-SRS-RNTI, C-RNTI, CS-RNTI(s), or SP-CSI-RNTI.
- *UE-specific* search space set configured by *SearchSpace* in *PDCCH-Config* with *searchSpaceType = ue-Specific* for a DCI format, the CRC of which is scrambled with C-RNTI, CS-RNTI(s), or SP-CSI-RNTI.

An example search space configuration for two devices is shown in Fig. 4.23. The UE determines a CORESET and PDCCH monitoring occasions for Type0-PDCCH common search space set, if it is not provided with RRC parameter *searchSpace-SIB1*. The Type0-PDCCH common search space set is defined by the CCE aggregation levels and the number of PDCCH candidates per CCE aggregation level. The CORESET configured for this search space set has CORESET index 0 and search space set index 0. If the UE is not provided with a CORESET for any of Type0A-PDCCH/Type1-PDCCH/Type2-PDCCH common search spaces, the corresponding CORESET would be the same as the CORESET for Type0-PDCCH common search space. The CCE aggregation levels and the number of PDCCH candidates per CCE aggregation level for Type0-PDCCH, Type0A-PDCCH, and Type2-PDCCH common search space are given in Table 4.5 [8].

The DM-RS antenna port associated with PDCCH reception in the Type0-PDCCH/Type0A-PDCCH/Type2-PDCCH common search spaces and for the corresponding PDSCH receptions as well as the DM-RS antenna port associated with SS/PBCH block reception are quasi-co-located with respect to delay spread, Doppler spread, Doppler shift, average delay,

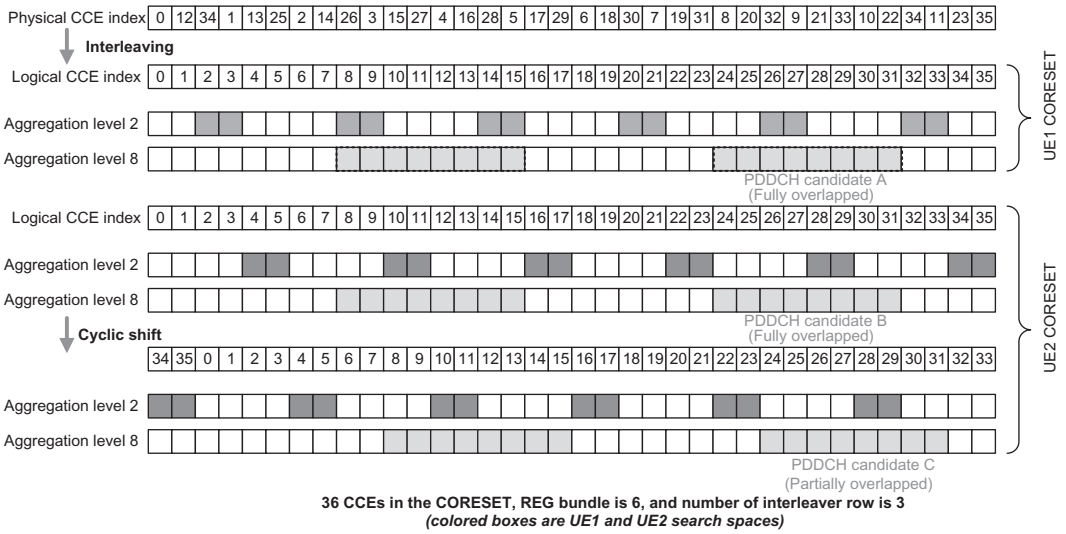


Figure 4.23
Example search space configuration for two devices [14].

Table 4.5: CCE aggregation levels and maximum number of PDCCH candidates per CCE aggregation level for Type0/ Type0A/Type2-PDCCH common search space [8].

CCE Aggregation Level	Number of Candidates
4	4
8	2
16	1

and spatial RX parameters. The value for the DM-RS scrambling sequence initialization is the cell ID. The subcarrier spacing and the cyclic prefix length for PDCCH reception with Type0A-PDCCH/Type1-PDCCH/Type2-PDCCH common search spaces are the same as for PDCCH reception with Type0-PDCCH common search space. The DM-RS antenna port associated with PDCCH reception and the associated PDSCH reception in Type1-PDCCH common search space are quasi-co-located with the SS/PBCH block identified in initial access procedure or with a received CSI-RS with respect to delay spread, Doppler spread, Doppler shift, average delay, and spatial RX parameters [8].

For each downlink BWP configured for a UE in the serving cell, the UE can be provided with $N_{CORESET} \leq 3$ CORESETs. For each CORESET, the RRC signaling provides the UE

with a CORESET index $0 \leq p < 12$; a DM-RS scrambling sequence initialization value; a precoder granularity for a number of REGs in the frequency domain where the UE can assume use of a same DM-RS precoder; a number of consecutive symbols; a set of resource blocks; CCE-to-REG mapping parameters; an antenna port QCL, from a set of antenna port QCLs, indicating QCL information of the DM-RS antenna port for PDCCH reception in a respective CORESET; and an indication for presence or absence of TCI field in DCI format 1_1 transmitted by PDCCH in the CORESET [8].

For each CORESET in a downlink BWP of a serving cell, the RRC parameter *frequencyDomainResources* provides a bitmap, whose bits have one-to-one correspondence with non-overlapping groups of six PRBs, in ascending order of the PRB index in the downlink BWP bandwidth of N_{RB}^{BWP} PRBs with starting position N_{start}^{BWP} where the first PRB of the first group of six PRBs is indexed as $6\lceil N_{start}^{BWP}/6 \rceil$. A group of six PRBs are allocated to a CORESET, if the corresponding bit value in the bitmap is set to one. If the UE receives the initial configuration of more than one TCI state through RRC parameter *TCI-States* but has not received a MAC CE activation command for at least one of the TCI states, the UE can assume that the DM-RS antenna port associated with PDCCH reception in the UE-specific search space is quasi-co-located with the SS/PBCH block that the UE has identified during the initial access procedure. If the UE has received a MAC CE activation command at least for one of the TCI states, it applies the activation command 3 ms after a slot where it transmits HARQ-ACK information for the PDSCH providing the activation command [8]. Table 4.6 provides the maximum number of PDCCH candidates $\max\left(M_{slot}^{PDCCH}(\mu)\right)$ across all CCE aggregation levels and across all DCI formats with different size in the same search space that the UE is expected to monitor per slot and per serving cell as a function of the subcarrier spacing. The table further provides the maximum number of non-overlapped CCEs that a UE is expected to monitor per slot and per serving cell as a function of the subcarrier spacing. The CCEs are considered non-overlapped, if they correspond to different CORESET indices or different first symbols for the reception of the respective PDCCH candidates [8].

Table 4.6: Maximum number of PDCCH candidates per slot and per serving cell as a function of subcarrier spacing [8].

μ	Maximum Number of Monitored PDCCH Candidates Per Slot and Serving Cell	Maximum Number of Non-overlapped CCEs Per Slot and Serving Cell
	$\max\left(M_{slot}^{PDCCH}(\mu)\right)$	$\max\left(C_{slot}^{PDCCH}(\mu)\right)$
0	44	56
1	36	56
2	22	48
3	20	32

For each downlink BWP that is configured for a UE in a serving cell, the UE is provided via RRC signaling with $S \leq 10$ search space sets. For each of those search space sets, the UE is provided with an search space set index $0 \leq s < 40$; an association between the search space set s and a CORESET p; a PDCCH monitoring periodicity of $k_{p,s}$ slots and a PDCCH monitoring offset of $\delta_{p,s}$ slots; a PDCCH monitoring pattern within a slot, indicating first symbol(s) of the CORESET within a slot for PDCCH monitoring; a number of PDCCH candidates $M_{p,s}^{L}$ per CCE aggregation level L; and an indication that search space set s is either a common or a UE-specific search space set via RRC signaling [8]. Alternative PDCCH mapping rules are illustrated in Fig. 4.24.

The UE can also be provided via RRC signaling with a time interval consisting of $T_{p,s} < k_{p,s}$ slots indicating a number of slots where the search space set s could exist. The information on the first symbol and the number of consecutive symbols for a CORESET, which results in a PDCCH candidate mapping to symbols of different slots, is not provided to the UE. The UE cannot assume that two PDCCH monitoring occasions, for the same search space set or for different search space sets, within the same CORESET are separated by a number of symbols that are less than the CORESET duration.

The UE determines the PDCCH monitoring occasion from the PDCCH monitoring periodicity, offset, and pattern within a slot. For search space set s in CORESET p, the UE determines the PDCCH monitoring occasion(s) in a slot with number n_{slot} in a frame with number n_{frame}, if $\left(n_{frame}N_{frame}^{slot} + n_{slot} - \delta_{p,s}\right) \bmod k_{p,s} = 0$. If the UE is informed in advance of the duration via RRC signaling, it would monitor PDCCH for search space set s in CORESET p for $T_{p,s}$ consecutive slots, starting from slot n_{slot} and would not monitor PDCCH for search space set s in CORESET p for the next $k_{p,s} - T_{p,s}$ consecutive slots [8].

A UE-specific search space at CCE aggregation level $L \in \{1, 2, 4, 8, 16\}$ is defined by a set of PDCCH candidates for CCE aggregation level L. For search space set s associated with CORESET p, the CCE indices for aggregation level L corresponding to PDCCH candidate $m_{s,n_{CI}}$ of the search space set in slot n_{slot} for an active BWP in the serving cell corresponding to carrier indicator field value n_{CI} are given as follows [8]:

$$\left\{ Y_{p,n_{slot}} + \left\lfloor \frac{m_{s,n_{CI}} N_{CCE,p}}{L \max\left(M_{p,s}^{(L)}\right)} \right\rfloor + n_{CI} \bmod \left\lfloor N_{CCE,p}/L \right\rfloor \right\} L + i, \quad \forall i = 0, \ldots, L-1$$

For common search spaces $Y_{p,n_{slot}} = 0$, whereas for a UE-specific search spaces $Y_{p,n_{slot}} = \left(A_p Y_{p,n_{slot}-1}\right) \bmod D$, $Y_{p,-1} = n_{RNTI}$, $A_p = 39827$, 39829, or 39839 if $p \bmod 3 = 0, 1$, or 2, respectively. In the preceding expression, $D = 65537$; $N_{CCE,p}$ is the number of CCEs, numbered from 0 to $N_{CCE,p} - 1$, in CORESET p; and n_{CI} denotes the carrier indicator field value, if the UE is configured via RRC signaling with a carrier indicator field in the serving cell on which the PDCCH is monitored; otherwise, including for any common search space $n_{CI} = 0$. Furthermore, $m_{s,n_{CI}} = 0, \ldots, M_{p,s,n_{CI}}^{(L)} - 1$, where $M_{p,s,n_{CI}}^{(L)}$ is the

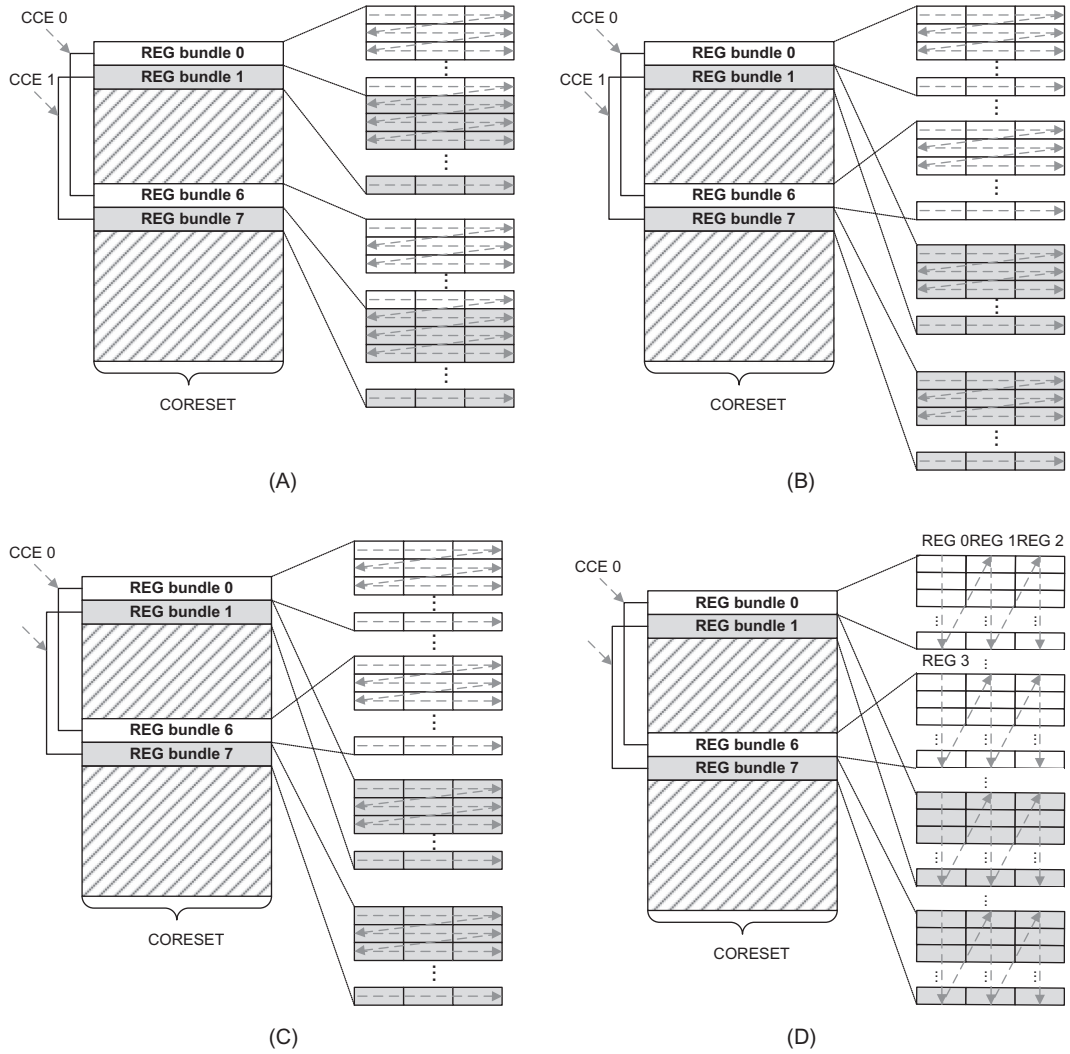


Figure 4.24
Illustration of alternative PDCCH mapping rules: (A) PDCCH candidate-level, (B) CCE-level, (C) REG bundle-level, and (D) REG-level.

number of PDCCH candidates that the UE is supposed to monitor at aggregation level L in a search space set s in a serving cell corresponding to n_{CI}. For any common search space, $\max(M^{(L)}_{p,s,n_{CI}}) = M^{(L)}_{p,s,0}$ whereas for a UE-specific search space, $\max(M^{(L)}_{p,s,n_{CI}})$ denotes the maximum of $M^{(L)}_{p,s,n_{CI}}$ over the configured values of n_{CI} for a CCE aggregation level L of search space set s in CORESET p; the RNTI value used for n_{RNTI} is the C-RNTI [8]. Example PDCCH search spaces at various aggregation levels are shown in Fig. 4.25.

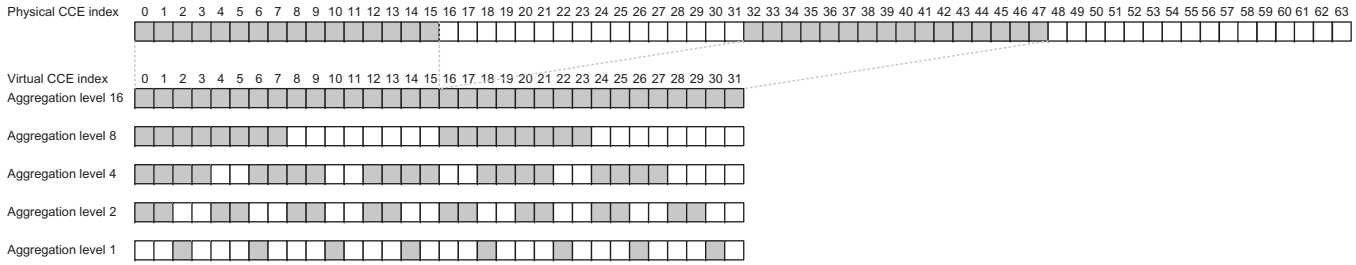


Figure 4.25
Example PDCCH search spaces at various aggregation levels.

When the total number of PDCCH candidates that a UE is configured to monitor in a slot exceeds the blind decoding limit of the UE, it must drop some of the candidates. While the number of blind decoding attempts can be controlled by the network through RRC configuration, the specification does not define a rule for dropping the PDCCH candidates, when the number of PDCCH candidates exceeds the blind decoding limit of the UE. As an example, the candidates can be prioritized according to the search space type and then according to search space set number within a search space type and finally according to aggregation level within a search space set. Once the blind decoding limit is reached, the remaining candidates can be discarded. If a UE has a limit of 56 CCEs and it is configured with a single CORESET in the slot that spans only one OFDM symbol then the CCE limit should not be an issue since the maximum number of PRBs on an RF carrier is 275, and this corresponds to less than 56 CCEs. However, if a UE is configured to monitor two-symbol or three-symbol CORESETs, or multiple CORESETs in a slot and the number of CCEs for all CORESETs is large, the CCE processing constraint can potentially limit the number of PDCCH candidates for which decoding can be attempted, considering that the CCE limit of 56 CCEs must be shared among many CORESETs in the slot.

4.1.3.2.5 Dynamic and Semi-persistent Scheduling

The MAC sublayer in a gNB includes dynamic resource schedulers that manage and allocate radio resources to active users in the downlink and uplink. Scheduling is performed in either dynamic or semi-static manner. Dynamic scheduling is the default mode-of-operation where the scheduler for each time interval, that is, a slot, determines which devices are going to transmit or receive and further configures the transmission parameters based on the measurement reports from the UEs. Since scheduling decisions are made frequently, it is possible to track fast variations of the user traffic as well as the channel quality, thus efficiently utilizing the available resources in order to maximize the network capacity. Semi-static scheduling implies that the transmission parameters are provided to the devices in advance and are not changed on a dynamic basis.

The scheduler operation takes into account the UE buffer status and the QoS requirements of each UE as well as the associated radio bearers when assigning radio resources among active UEs (see Fig. 4.26). The schedulers assign network resources to the UEs by considering the radio conditions as seen by the UEs, identified through measurements made at the gNB and/or reported by the UE. The schedulers assign radio resources in a unit of slot, for example, one mini-slot, one slot, or multiple slots and resource assignments consist of radio resources (time, frequency, code, space, power). The UEs identify the allocated resources by receiving a scheduling decision (resource assignment) through PDCCH. The UE periodically or on-demand basis conducts measurements to support scheduler operation. The uplink buffer status reports (measuring the data that is buffered in the logical channel queues in the UE) are used to provide support for QoS-aware packet scheduling. Power headroom reports

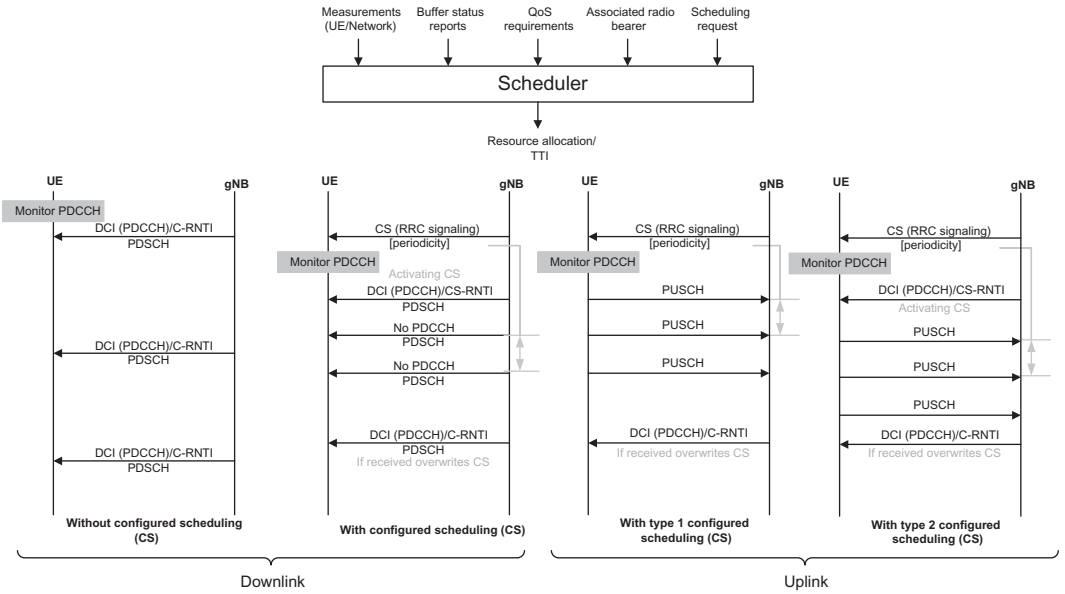


Figure 4.26
Illustration of downlink and uplink dynamic and configured scheduling [11].

(measuring the difference between the nominal UE maximum transmit power and the estimated power for uplink transmission) are used to provide support for power aware packet scheduling [11].

In the downlink, the gNB dynamically allocates radio resources to the active UEs using their respective C-RNTIs scrambled into their PDCCH(s) CRCs to uniquely identify the users. A UE always monitors certain PDCCH(s) candidates in order to find possible downlink/uplink assignments where the process is coordinated with the UE-specific DRX cycles, when configured. When carrier aggregation is configured for the UE, the same C-RNTI applies to all serving cells. In addition, with SPS, the gNB can allocate downlink resources for the initial HARQ transmissions to UEs. The RRC signaling defines the periodicity of the configured downlink assignments while PDCCH scrambled with CS-RNTI can either signal and activate the configured downlink assignment, or deactivate it, that is, a PDCCH addressed to a UE using its CS-RNTI indicates that the downlink assignment can be implicitly reused according to the periodicity defined by the RRC signaling, until deactivated. When a configured downlink assignment is active, if the UE cannot find its C-RNTI on the PDCCH(s), a downlink transmission according to the configured downlink assignment is assumed; otherwise, if the UE finds its C-RNTI on the PDCCH(s), the PDCCH allocation overrides the configured downlink assignment. When carrier aggregation is configured, one configured downlink assignment can be signaled per serving cell. When bandwidth

adaptation is configured, one configured downlink assignment can be signaled per BWP. On each serving cell, there can be only one configured downlink assignment active at a time, and multiple configured downlink assignment can be simultaneously active on different serving cells. Activation and deactivation of configured downlink assignments are independent among the serving cells [11].

In the uplink, the gNB can dynamically allocate resources to the UEs by scrambling their respective C-RNTI with PDCCH(s) CRCs. A UE always monitors the PDCCH(s) in order to find possible grants for uplink transmission when its downlink reception is enabled where the activity is synchronized with the UE DRX cycles. When carrier aggregation is configured, the same C-RNTI applies to all serving cells. In addition, with configured grants, the gNB can allocate uplink resources for the initial HARQ transmissions to the UEs. Two types of configured uplink grants are defined in NR: Type 1, where the RRC signaling directly provides the configured uplink grant (including the periodicity); and Type 2, where the RRC signaling defines the periodicity of the configured uplink grant while PDCCH addressed to the UE using its CS-RNTI can either signal and activate the configured uplink grant, or deactivate it, that is, a PDCCH addressed to the UE using its CS-RNTI would indicate that the uplink grant can be implicitly reused according to the periodicity defined by RRC, until deactivated (see Fig. 4.26). When a configured uplink grant is active, if the UE cannot find its C-RNTI/CS-RNTI on the PDCCH(s), an uplink transmission according to the configured uplink grant can be attempted. Otherwise, if the UE finds its C-RNTI/CS-RNTI on the PDCCH(s), the PDCCH allocation overrides the configured uplink grant. Retransmissions other than repetitions are explicitly allocated via PDCCH(s). When carrier aggregation is configured, one configured uplink grant can be signaled per serving cell. Similarly, when bandwidth adaptation is configured, one configured uplink grant can be signaled per BWP. In each serving cell, there can be only one configured uplink grant active at a time. A configured uplink-grant for one serving cell can either be of Type 1 or Type 2. For Type 2, activation and deactivation of configured uplink grants are independent among the serving cells. When SUL is configured, a configured uplink grant can only be signaled for one of the two uplink carriers of the cell [11].

4.1.4 Synchronization Signals

In order to connect/attach to the network, a UE must perform initial cell search and downlink synchronization. The objective of initial cell search is to find a strong cell signal for connection establishment, to obtain an estimate of frame timing, to obtain cell identification, and to find the reference signals for coherent demodulation of PBCH and PDCCH. For this purpose, the PSS and SSS are used. The PSS and SSS are transmitted in SSBs together with PBCH. The blocks are transmitted per slot at a fixed slot location. During initial cell search, the UE correlates the received signals and the synchronization signal sequences by means

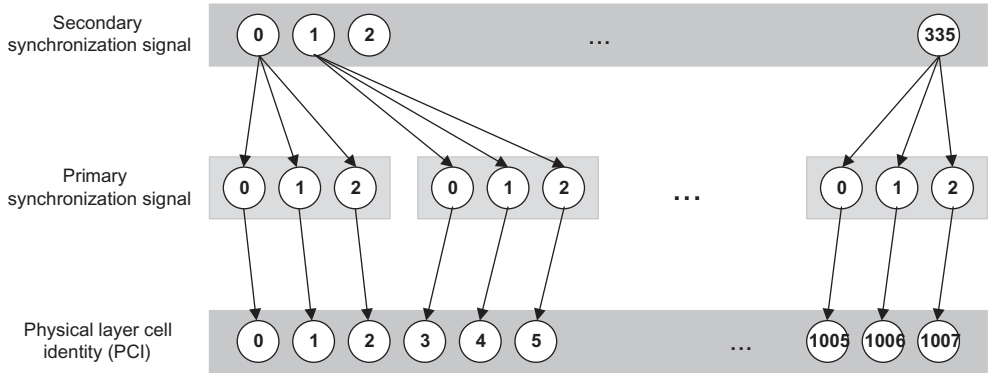


Figure 4.27
Derivation of PCI based on PSS and SSS sequences [6].

of matched filtering and attempts to locate the PSS in order to obtain symbol and half-frame timing. It then attempts to find the SSS in order to detect the cyclic prefix length as well as the duplexing scheme and to obtain the exact frame timing based on matched filter results for the PSS and SSS. It then proceeds to detect the cell identity from the reference signals sequence index and to decode the PBCH for the purpose of obtaining the minimum SI. The synchronization signal (SS) blocks are organized into SS bursts and SS bursts are organized into SS burst sets that are periodically transmitted in order to support beamforming operation.

In NR, there are 1008 unique physical-layer cell identities, that is, an increased number compared to 504 in LTE in order to provide sufficient deployment flexibility in dense network topologies. As shown in Fig. 4.27, the NR physical-layer cell identities are in 336 unique physical-layer cell-identity groups, each group containing three unique identities. Each NR cell ID can be jointly represented by a PSS/SSS combination. The PSS consists of three frequency-domain binary BPSK length-127 M-sequences,[6] and the SSS corresponds to 336 Gold sequences with length-127. Both of these signals are mapped into 127 contiguous subcarriers. A physical-layer cell identity is uniquely defined by a number in the range of 0−335, representing the physical-layer cell-identity group, and a number in the range of 0−2, representing the physical-layer identity within the physical-layer cell-identity group as

[6] Maximum length sequences are pseudo-random binary sequences that are generated using maximal linear feedback shift registers. The M-sequences are periodic and reproduce every binary sequence that can be reproduced by the shift registers (i.e., for length-m registers, they produce a sequence of length $2^m - 1$). An M-sequence is spectrally flat with the exception of a near-zero DC term. Since M-sequences are periodic and shift registers cycle through every possible binary value with the exception of the zero vectors, the registers can be initialized to any state with the exception of the zero vectors. A binary polynomial over GF(2) can be associated with the linear feedback shift register. The degree of polynomial is equal to the length of the shift register and the coefficients that are either 0 or 1 correspond to the taps of the register.

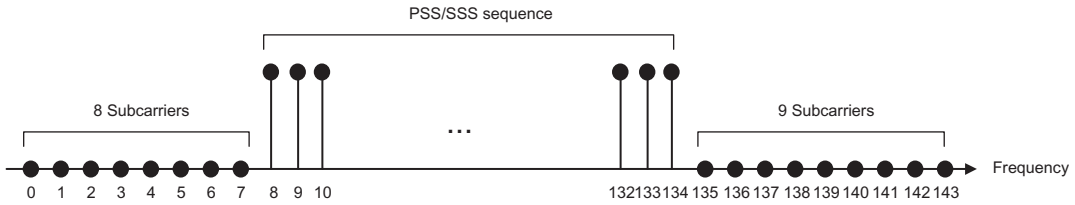


Figure 4.28
Illustration of PSS/SSS sequence mapping to the resource elements [6].

in the following formula $N_{ID}^{cell} = 3N_{ID}^{(1)} + N_{ID}^{(2)}$ where $N_{ID}^{(1)} \in \{0, 1, \ldots, 335\}$ and $N_{ID}^{(2)} = \{0, 1, 2\}$. The cell number information is carried in PSS, whereas the cell group number is carried in SSS. It is worthwhile mentioning that the physical beams associated with an SS blocks are transparent to the UE, since the latter only sees the equivalent synchronization signals and the PBCH after precoding and/or beamforming operations that are implementation specific.

4.1.4.1 Primary Synchronization Sequence

The PSS sequence $d_{PSS}(n)$ is defined by $d_{PSS}(n) = 1 - 2x(m)$ where $m = \left(n + 43N_{ID}^{(2)}\right) \bmod 127$ and $0 \leq n < 127$. In the latter expression, $x(i+7) = [x(i+4) + x(i)] \bmod 2$ and $\begin{bmatrix} x(6) & x(5) & x(4) & x(3) & x(2) & x(1) & x(0) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}$. The sequence of symbols $d_{PSS}(0), \ldots, d_{PSS}(126)$ containing the PSS is scaled by a factor of β_{PSS} in order to adjust its transmission power and is mapped to resource elements (k, l) in increasing order of k where k and l represent the frequency and time indices, respectively (see Table 4.3), within one SS/PBCH block. The PSS sequence mapping to resource elements in the frequency domain is illustrated in Fig. 4.28. Furthermore, the time-frequency structure and timing of the PSS transmission are depicted in Fig. 4.29.

4.1.4.2 Secondary Synchronization Sequence

The secondary synchronization signal sequence $d_{SSS}(n)$ is defined as $d_{SSS}(n) = [1 - 2x_0((n + m_0) \bmod 127)][1 - 2x_1((n + m_1) \bmod 127)]$ where $m_0 = 15\left\lfloor N_{ID}^{(1)}/112 \right\rfloor + 5N_{ID}^{(2)}$, $m_1 = N_{ID}^{(1)} \bmod 112$ and $0 \leq n < 127$. In the latter expression, $x_0(i+7) = \left[x_0(i+4) + x_0(i)\right] \bmod 2$ and $x_1(i+7) = \left[x_1(i+1) + x_1(i)\right] \bmod 2$ where $\begin{bmatrix} x_0(6) & x_0(5) & x_0(4) & x_0(3) & x_0(2) & x_0(1) & x_0(0) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} x_1(6) & x_1(5) & x_1(4) & x_1(3) & x_1(2) & x_1(1) & x_1(0) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$. The sequence of symbols $d_{SSS}(0), \ldots, d_{SSS}(126)$ containing the secondary synchronization signal is scaled by a factor of β_{SSS} and is mapped to resource elements (k, l) in increasing order of k where k and l represent the frequency and time indices, respectively (see Table 4.3),

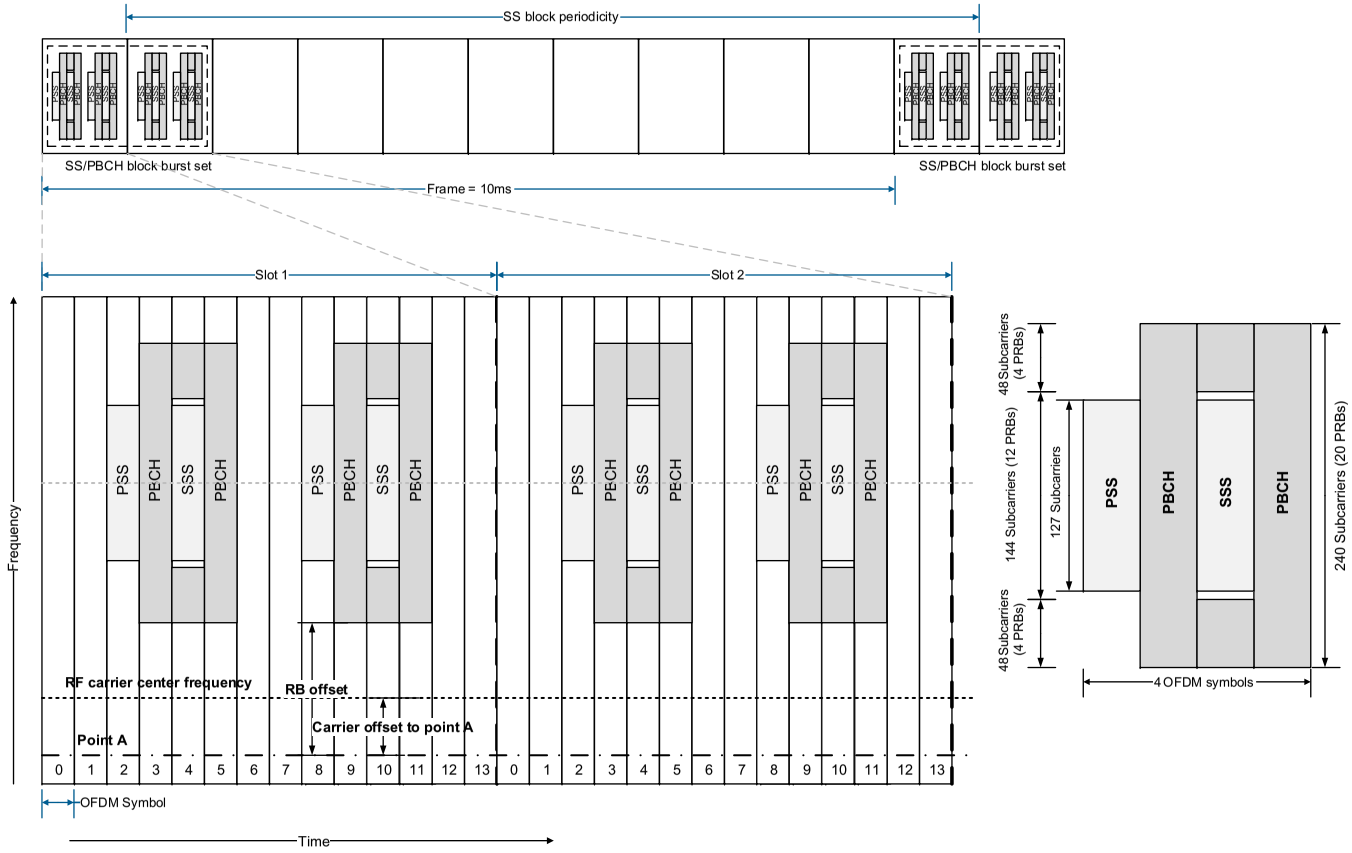


Figure 4.29
Time-frequency structure of the PSS within an SS block [6,56].

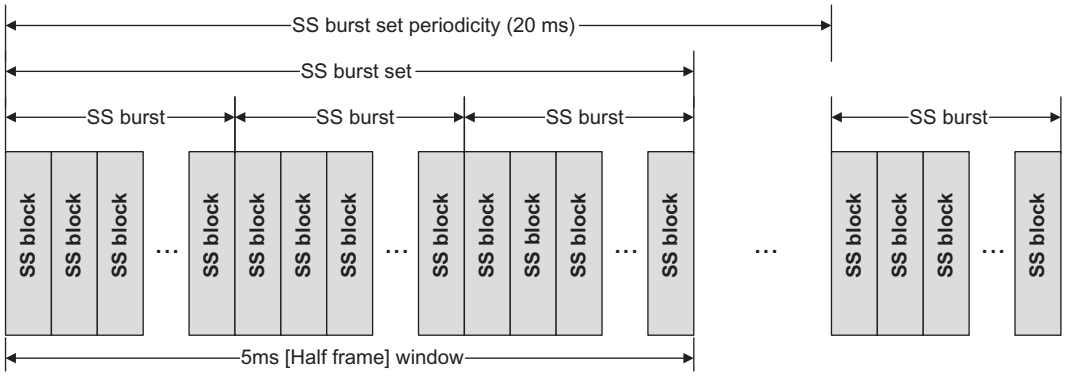


Figure 4.30
SS block structure and timing [6,8].

within one SS/PBCH block [6]. The SSS sequence mapping to resource elements in the frequency domain is illustrated in Fig. 4.28. Furthermore, the time-frequency structure and timing of the SSS transmission are depicted in Fig. 4.29.

4.1.4.3 Synchronization Signal Blocks

In NR, the primary and secondary synchronization signals are used by the UE for initial cell search and to obtain frame timing, Cell ID, and to find the reference signals for coherent demodulation of other channels. The PSS, SSS, and PBCH are time-multiplexed and transmitted in an SSB with the same numerology. One or more SS block(s) constitute an SS burst, and one or more SS bursts form an SS burst set as illustrated in Fig. 4.30. The SS burst sets are transmitted periodically. An SS block consists of four consecutive OFDM symbols. Regardless of the SS burst set composition, the transmission of SSBs within an SS burst set is confined to a 5 ms window to help the UEs reduce power consumption and complexity for radio resource management-related measurements. Fig. 4.31 compares the synchronization signals and broadcast channel transmission timings of LTE and NR.

The SS block is transmitted periodically with a period which may be configured between 5 and 160 ms. However, the UEs performing initial cell search or handover can assume that the SS block is repeated every 20 ms. Each SS burst set is always confined to a 5 ms window located either in the first or the second half of a 10 ms radio frame. This allows a UE that is searching for an SS block in the frequency domain to know the time duration it should pause at each frequency before retuning to the next frequency within the synchronization raster, concluding that there is no PSS/SSS present at that frequency. The 20 ms SS block periodicity is four times longer than the corresponding 5 ms periodicity of LTE PSS/SSS transmission (see Fig. 4.31). The longer SS block period was selected to improve the NR network energy efficiency and to reduce the layer-1 overhead. The disadvantage of a

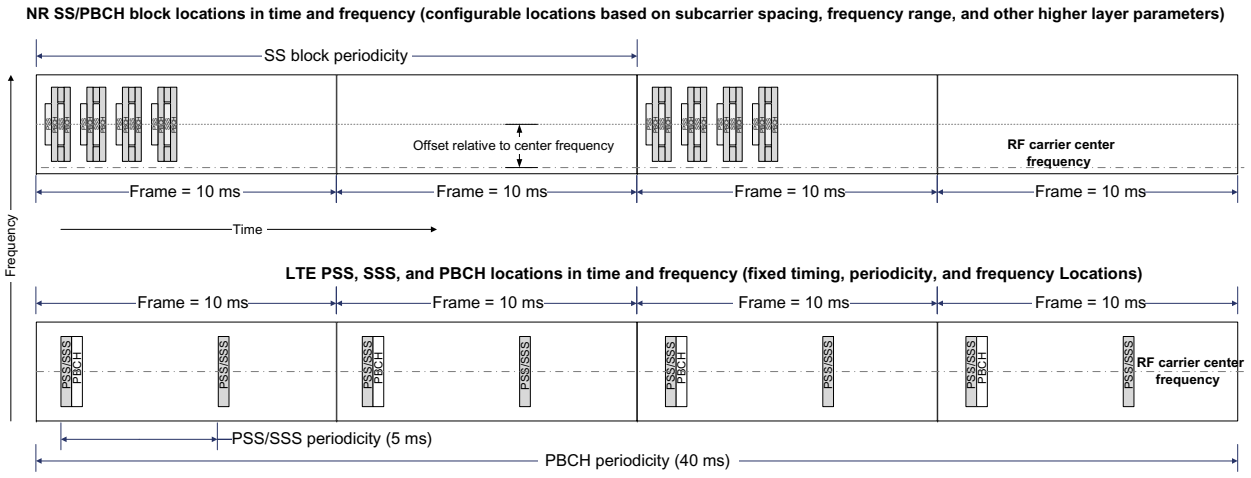


Figure 4.31
Comparison of LTE and NR synchronization and broadcast channel transmission timing.

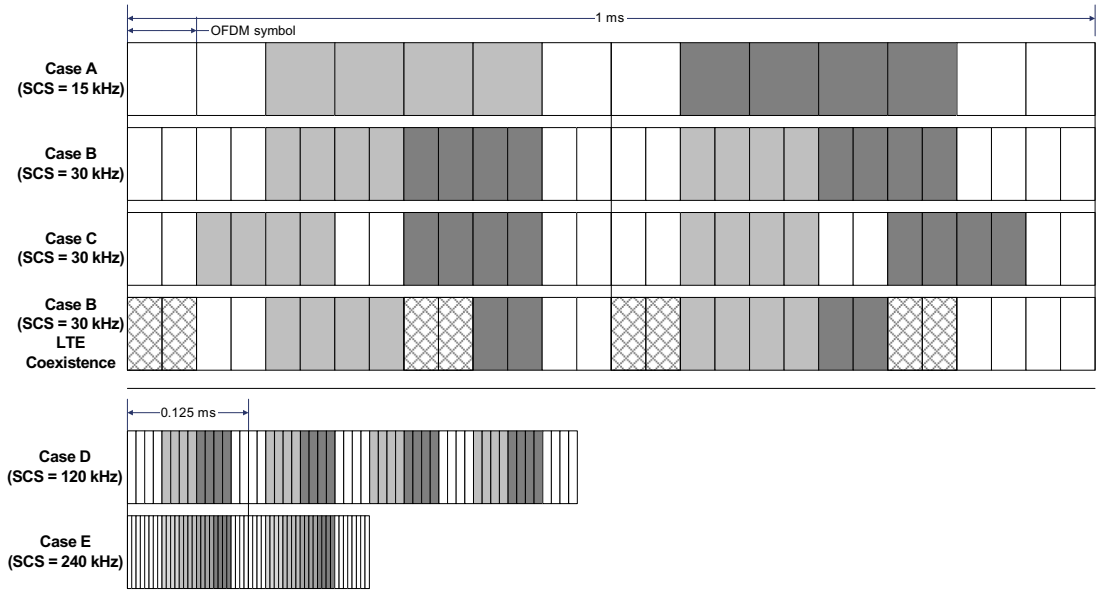


Figure 4.32
Structure and timing of SS/PBCH block transmission with various numerologies [6,31].

longer SS block period is that a device must pause at each frequency for a longer time in order to conclude that there is no PSS/SSS at the frequency. However, this is compensated by the sparse synchronization raster relative to LTE, which reduces the number of frequency-domain locations at which a device must search for an SS block.

The maximum number of SSBs within an SS burst set is 4 for frequency ranges up to 3 GHz, 8 for 3–6 GHz, and 64 for above 6 GHz in order to achieve a trade-off between coverage and layer-1 overhead. Furthermore, the number of actually transmitted SSBs could be less than the maximum number. The position(s) of actually transmitted SSBs can be signaled to the UEs in order to assist their RRC_CONNECTED or RRC_IDLE mode measurements and to help the UEs in RRC_CONNECTED and potentially RRC_IDLE mode to receive downlink data/control in unused SSBs. The structure and timing of SSB transmission with various numerologies is illustrated in Fig. 4.32, where a number of symbols are reserved for downlink control at the beginning of the slot, and some symbols are reserved for guard period and uplink control to allow UL/DL switching and fast uplink feedback. The SSB pattern corresponding to 15 and 30 kHz subcarrier spacing can provide more UL/DL switching opportunities in TDD mode. The SSB pattern for 30 kHz subcarrier spacing can be used to facilitate LTE-NR coexistence in the downlink in an FDD system, considering the locations of LTE PDCCH and cell-specific reference signals in the symbols with LTE default 15 kHz subcarrier spacing.

Within a PBCH transmission time interval update period of 80 ms, there are 16 possible positions for an SS burst set, if we consider the minimum period for an SS burst set to be 5 ms. The 16 possible positions of an SS burst set can be identified by the three least significant bits of the SFN and one-bit half radio frame index. The SSBs can be repeated within an SS burst set. When the UE detects an SSB, it will acquire the timing information from its PBCH, from which the UE is able to identify the radio frame number, the slot index in a radio frame, and OFDM symbol index in a slot. The timing information includes 10 bits for SFN, 1 bit for half radio frame index, and 2, 3, or 6 bits for SSB time index for frequency ranges up to 3, 3−6, and 6−52.6 GHz, respectively. Within the SSB indices, two or three LSBs are carried by changing the DM-RS sequence of PBCH. Thus, for the sub-6 GHz frequency range, the UE can acquire the SSB index without decoding the PBCH. It also facilitates PBCH soft combining over multiple SSBs as these SSBs with different indices carry the same PBCH payload [72]. As shown in Fig. 4.32, for a half frame with SS/PBCH blocks, the first symbol indices of the candidate SS/PBCH blocks are determined according to the subcarrier spacing of SS/PBCH blocks as follows, where index 0 corresponds to the first symbol of the first slot in a half-frame [8]:

- *Case A (SCS = 15 kHz):* The first symbols of the candidate SS/PBCH blocks have indices of $\{2, 8\} + 14n$ where $n = 0,\ 1$ for carrier frequencies $f_c \leq 3$ GHz and $n = 0, 1, 2, 3$ for carrier frequencies 3 GHz $\leq f_c \leq 6$ GHz.
- *Case B (SCS = 30 kHz):* The first symbols of the candidate SS/PBCH blocks have indices $\{4, 8, 16, 20\} + 28n$ where $n = 0$ for carrier frequencies $f_c \leq 3$ GHz, $n = 0,\ 1$ for carrier frequencies 3 GHz $\leq f_c \leq 6$ GHz.
- *Case C (SCS = 30 kHz):* The first symbols of the candidate SS/PBCH blocks have indices $\{2, 8\} + 14n$ where $n = 0,\ 1$ for paired spectrum operation and carrier frequencies $f_c \leq 3$GHz and $n = 0, 1, 2, 3$ for carrier frequencies 3 GHz $\leq f_c \leq 6$ GHz. For unpaired spectrum operation, $n = 0,\ 1$ for carrier frequencies $f_c \leq 2.4$ GHz, $n = 0, 1, 2, 3$ for carrier frequencies 2.4 GHz $\leq f_c \leq 6$ GHz.
- *Case D (SCS = 120 kHz):* The first symbols of the candidate SS/PBCH blocks have indices $\{4, 8, 16, 20\} + 28n$ where $n = 0, 1, 2, 3, 5, 6, 7, 8, 10, 11, 12, 13, 15, 16, 17, 18$ for carrier frequencies $f_c \geq 6$ GHz.
- *Case E (SCS = 240 kHz):* The first symbols of the candidate SS/PBCH blocks have indices $\{8, 12, 16, 20, 32, 36, 40, 44\} + 56n$ where $n = 0, 1, 2, 3, 5, 6, 7, 8$ for carrier frequencies $f_c \geq 6$ GHz.

In order to support multi-beam operation, particularly in high-frequency scenarios, NR introduced the SSB, which comprises primary and secondary synchronization signals and PBCH. As illustrated in Fig. 4.33, a given SSB is repeated within an SS burst set, which is potentially used for gNB beam-sweeping transmission. The SS burst set is confined to a 5 ms window and transmitted periodically. For initial cell selection, the UE assumes a default SS burst set periodicity of 20 ms. The main advantage of SS burst set is that

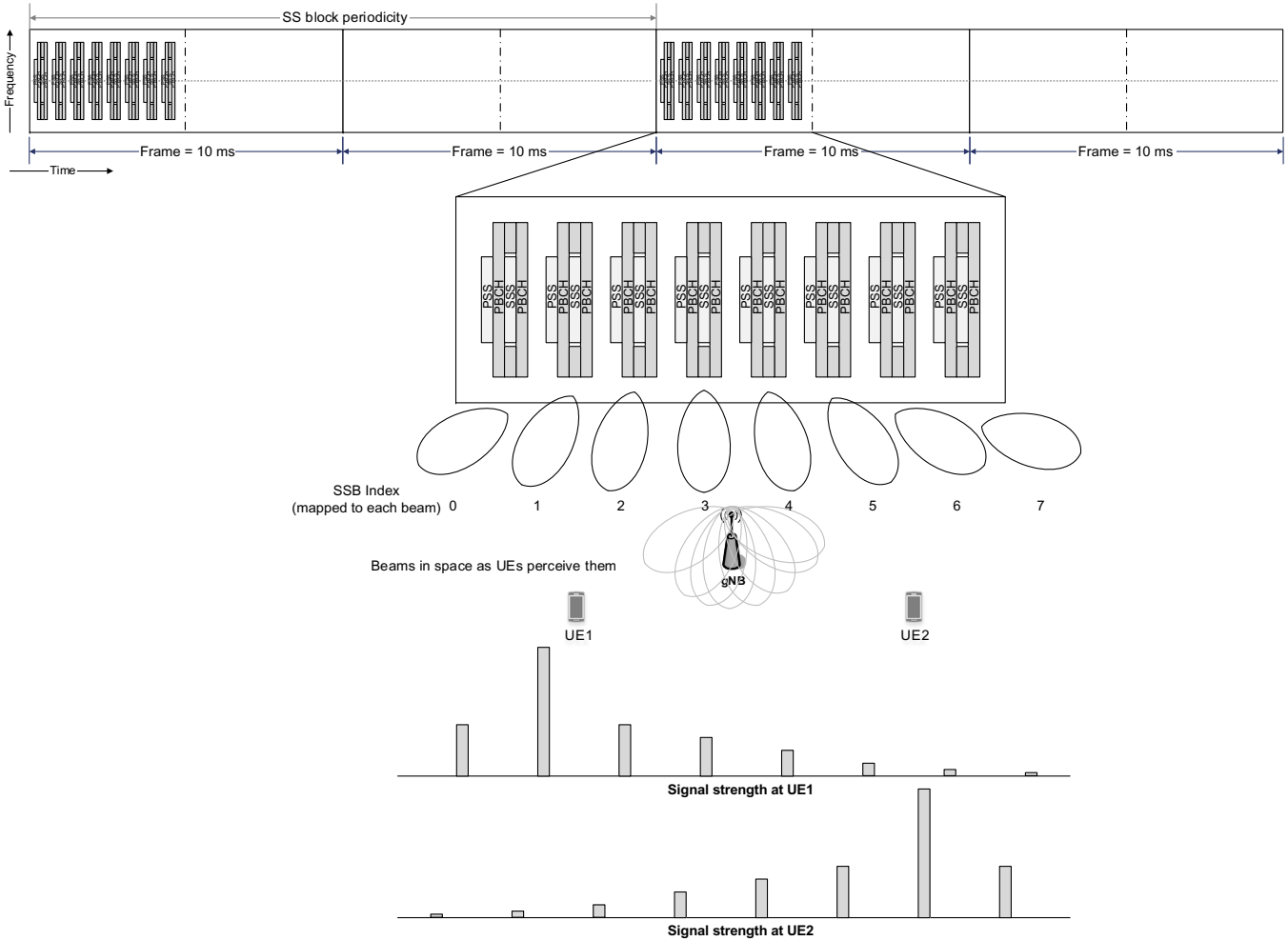


Figure 4.33
Example beam sweeping and correspondence to the transmission of SS burst set [30].

time-division multiplexing beam-sweeping allows for low-cost analog antenna array architectures. Frequency-division multiplexing is another approach that could have been potentially adopted in NR; however, it would have precluded use of analog antenna array architectures. This feature, although particularly useful for mmWave operation, can also be leveraged at lower frequency bands.

In LTE, the PSS/SSS and PBCH are always located at the center of the RF carrier. Thus once an LTE device detects the PSS/SSS, it has already found the center frequency of the carrier. The drawback of this approach is that a device with no a priori knowledge of the frequency-domain carrier position must search for PSS/SSS at all possible carrier positions. To allow a faster cell search in NR, the possible SS block locations for each frequency band are a limited set of frequencies referred to as the synchronization raster. Therefore, instead of searching for an SS block at each carrier raster, a UE only needs to search for an SS block within the sparse set of synchronization raster. Since NR carriers can still be located at an arbitrary position on the carrier raster, the SS block may not be necessarily located at the center of a carrier, and it may not be aligned with the resource block grid due to different numerologies. Thus once the SS block has been detected, the device must be explicitly informed about the exact SS block frequency-domain position on the carrier. This is achieved by means of information partly within PBCH and partly within the RMSI.

As we mentioned earlier, an SS/PBCH block consists of four OFDM symbols in the time domain, numbered in increasing order from 0 to 3 within the SS/PBCH block, where PSS, SSS, and PBCH with the associated DM-RS are mapped to symbols as shown in Fig. 4.29. In the frequency domain, an SS/PBCH block consists of 240 contiguous subcarriers with the subcarriers numbered in increasing order from 0 to 239 within the SS/PBCH block. There are two types of SS/PBCH block, that is, Type A and Type B, where the former is specified for operation in sub-6 GHz frequency range and the latter is defined for mmWave bands. The frequency-domain location of SS/PBCH block is defined by parameter k_{SSB} which provides the subcarrier offset from subcarrier 0 in common resource block N_{CRB}^{SSB} to subcarrier 0 of the SS/PBCH block. The common resource block parameter N_{CRB}^{SSB} is derived from the RRC parameter *offsetToPointA*. The four LSBs of k_{SSB} are derived from the RRC parameter *ssb-SubcarrierOffset* where, for SS/PBCH block Type A, the most significant bit of k_{SSB} is given by $a_{N_{MIB}+5}$ in the PBCH payload [6,7]. If *ssb-SubcarrierOffset* is not provided, k_{SSB} is derived from the frequency difference between the SS/PBCH block and Point A. The complex-valued symbols corresponding to resource elements that are part of a common resource block partially or fully overlap with an SS/PBCH block and are not used for SS/PBCH transmission. For an SS/PBCH block, a single-antenna port and the same cyclic prefix length and subcarrier spacing are used for transmission of PSS, SSS, PBCH, and DM-RS for PBCH. For SS/PBCH block Type A, $\mu \in \{0, 1\}$ and $k_{SSB} \in \{0, 1, 2, \ldots, 23\}$ with the quantities k_{SSB} and N_{CRB}^{SSB} expressed in terms of 15 kHz subcarrier spacing. For

SS/PBCH block Type B, $\mu \in \{3, 4\}$ and $k_{SSB} \in \{0, 1, 2, \ldots, 11\}$ where the quantity k_{SSB} expressed in terms of the subcarrier spacing provided by the RRC parameter *subCarrierSpacingCommon*, and N_{CRB}^{SSB} is defined in terms of 60 kHz subcarrier spacing. The center of subcarrier 0 of resource block N_{CRB}^{SSB} coincides with the center of subcarrier 0 of a common resource block with the subcarrier spacing provided by the RRC parameter *subCarrierSpacingCommon*. This common resource block overlaps with subcarrier 0 of the first resource block of the SS/PBCH block. The SS/PBCH blocks are transmitted with the same block index on the same center frequency location which are quasi-co-located with respect to Doppler spread, Doppler shift, average gain, average delay, delay spread, and spatial RX parameters (when applicable) [6].

4.1.5 Physical Downlink Shared Channel

The downlink physical layer processing of transport channels consists of several steps as shown in Fig. 4.34 including CRC calculation and attachment to the TBs where a 24-bit CRC for payloads larger than 3824 bits or otherwise 16-bit CRC is attached; code block segmentation and code block CRC attachment; channel coding based on LDPC codes; physical-layer HARQ processing and rate matching; bit-interleaving; modulation; layer mapping and precoding; and mapping to assigned resources and antenna ports. At least one symbol with DM-RSs is present on each layer in which PDSCH is transmitted to a UE. The number of DM-RS symbols and RE mapping is configured by the RRC parameters. The PT-RS may be transmitted on additional symbols to aid receiver phase tracking.

As shown in Fig. 4.34, in the first stage of PDSCH processing, the entire TB $a_0, a_1, \ldots, a_{N_{TB}-1}$ is used to calculate the CRC parity bits $p_0, p_1, \ldots, p_{L_{CRC}-1}$ where N_{TB} and L_{CRC} denote the (TB) payload size and the number of CRC parity bits, respectively. The number of parity bits is set to 24 and the CRC generator polynomial $g_{CRC24A}(D) = [D^{24} + D^{23} + D^{18} + D^{17} + D^{14} + D^{11} + D^{10} + D^7 + D^6 + D^5 + D^4 + D^3 + D + 1]$ is used, however, if $N_{TB} \geq 3824$, L_{CRC} is set to 16 bits and the generator polynomial $g_{CRC16}(D) = [D^{16} + D^{12} + D^5 + 1]$ is used. The output bits following the CRC attachment are denoted by $b_0, b_1, \ldots, b_{B-1}$ where $B = N_{TB} + L_{CRC}$. For initial transmission of a TB with coding rate R, which is determined by the MCS index contained in the DCI, and the retransmissions of the same TB, each code block of the TB is encoded with LDPC base graph 2 (see the section on channel coding), if $N_{TB} \leq 292$, if $N_{TB} \leq 3824$ and $R \leq 0.67$ or if $R \leq 0.25$; otherwise, LDPC base graph 1 is used as depicted in Fig. 4.35.

As shown in Fig. 4.35, the maximum size of a code block K_{cb} is 8448 bits for LDPC base graph 1, 3840 for LDPC base graph 2. The code blocks whose size exceeds these limits would be segmented and appended with an additional CRC of length $L_{CRC} = 24$ bits. The input bits to the code block segmentation denoted as $b_0, b_1, \ldots, b_{B-1}$, where B is the number of bits in the TB (including the CRC), are then processed by code block segmentation

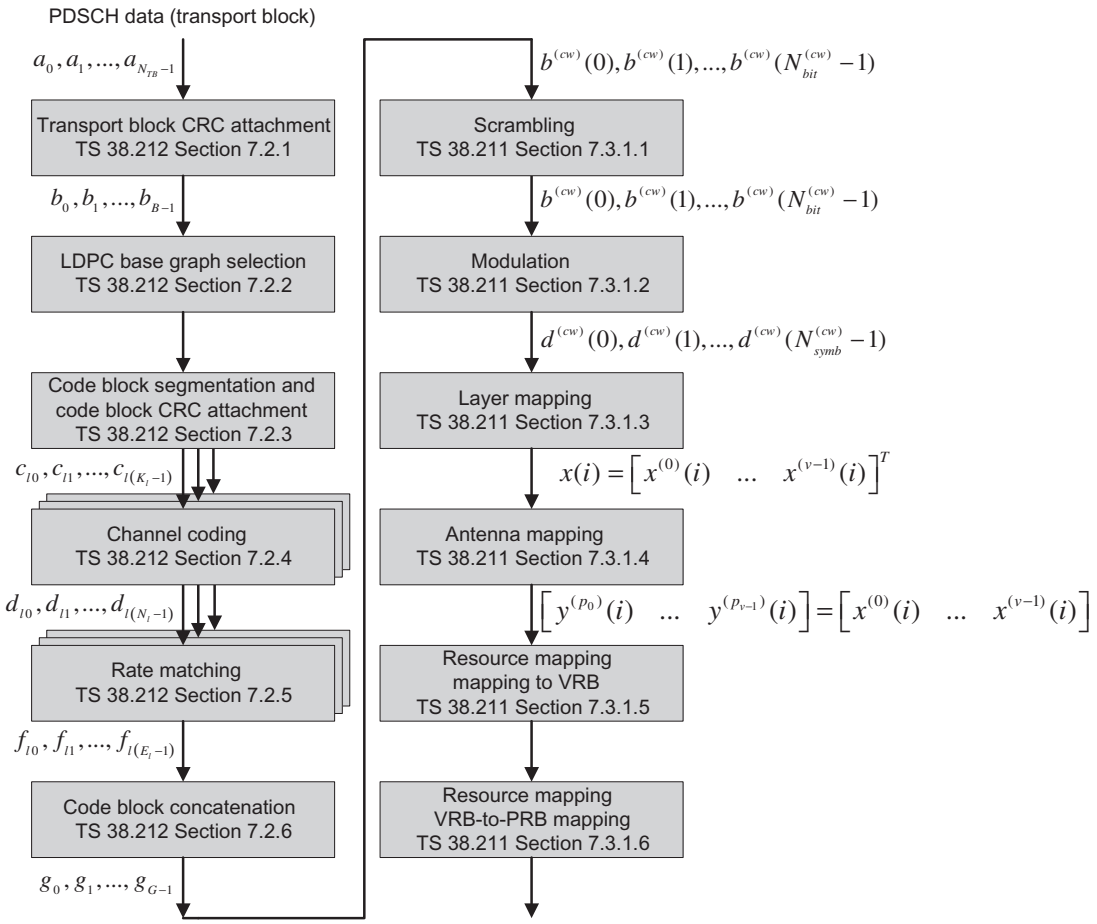


Figure 4.34
Physical layer processing of PDSCH [6,30].

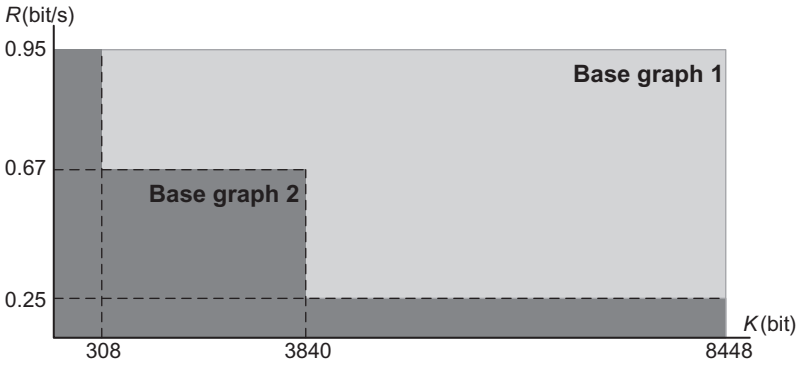


Figure 4.35
Usage of NR LDPC base graphs [55].

Table 4.7: NR low-density parity check code lifting factors Z_c [7].

Set Index i_{LS}	Set of Lifting Sizes Z_C
0	{2,4,8,16,32,64,128,256}
1	{3,6,12,24,48,96,192,384}
2	{5,10,20,40,80,160,320}
3	{7,14,28,56,112,224}
4	{9,18,36,72,144,288}
5	{11,22,44,88,176,352}
6	{13,26,52,104,208}
7	{15,30,60,120,240}

Table 4.8: NR low-density parity check base graphs parameters [7].

Parameter	Base Graph 1	Base Graph 2
Minimum code rate $R_{\min}$	1/3	1/5
Base matrix size	46×68	42×52
Number of systematic columns K_b	22	10
Maximum information block size K_{cb}	8448($= 22 \times 384$)	3840($= 10 \times 384$)
Number of non-zero elements	316	197

followed by code block CRC attachment, resulting in the output bits $c_{l0}, c_{l1}, \ldots, c_{l(K_l-1)}$ where index $0 \leq l < C$ represents the code block number and K_l denotes the number of bits for lth code block. The total number of code blocks C is determined by $C = \lceil B/(K_{cb} - L_{CRC}) \rceil$. The code blocks are then fed into the channel coding unit. The LDPC encoded bits are denoted by $d_{l0}, d_{l1}, \ldots, d_{l(N_l-1)}$ where the value of N_l is calculated as follows: If the bit sequence input for a given code block to channel coding is denoted by $c_0, c_1, \ldots, c_{K-1}$ where K is the number of bits to encode, and the LDPC encoded bits are denoted by $d_0, d_1, \ldots, d_{N-1}$ then $N = 66Z_c$ for LDPC base graph 1 and $N = 50Z_c$ for LDPC base graph 2, where the lifting factor Z_c is given in Table 4.7 [7]. The NR LDPC base graphs parameters are shown in Table 4.8.

The rate matching for LDPC code is performed on code block basis and consists of bit selection and bit-level interleaving. The input bit sequence to rate matching block is denoted as $d_0, d_1, \ldots, d_{N-1}$ which is written into a circular buffer of length N_{cb} for code block l, where code length N was defined earlier. Let us assume $N_{cb} = N$ for the lth code block, if $I_{LBRM} = 0$[7] and in other cases $N_{cb} = \min(N, N_{REF})$, in which $N_{REF} = \lfloor TBS_{LBRM}/(R_{LBRM}C) \rfloor$, C is the number of code blocks of the transport block, $R_{LBRM} = 2/3$ and TBS_{LBRM} for DL-SCH/PCH is obtained from Table 4.16, taking into consideration the maximum number of layers for one TB supported by the UE in the serving cell; the maximum modulation order

[7] Limited-buffer rate matching (LBRM) is a technique to process HARQ with reduced requirements for soft buffer sizes while maintaining the peak data rates. LBRM shortens the length of the virtual circular buffer of the code block segments for certain large sizes of transport blocks, thus sets a lower bound on the code rate.

configured for the serving cell, if configured by higher layers; otherwise, a maximum modulation order of $Q_m = 6$ is assumed for DL-SCH; and the maximum coding rate of 948/1024. Due to unequal amplitude of demodulated log likelihood ratios (LLRs) for 16QAM/64QAM/256QAM modulated symbols, it is necessary to consider a bit interleaving scheme for high-order modulations (see Fig. 4.37) in order to enhance the performance of the LDPC codes. The output bit sequence of the bit-interleaving function is the input to code block concatenation. The code block concatenation consists of sequentially concatenating the rate-matched outputs of different code blocks. The output bit sequence of code block concatenation is denoted by $g_0, g_1, \ldots, g_{G-1}$ where G is the total number of coded bits for transmission [7].

Rate matching is performed separately for each code block by puncturing a fraction of systematic bits. Depending on the code block size, the fraction of punctured systematic bits can be up to one-third of the systematic bits. The remaining coded bits are written into a circular buffer, starting with the non-punctured systematic bits and continuing with parity bits as shown in Fig. 4.36. The selection of the bits for transmission is based on reading the

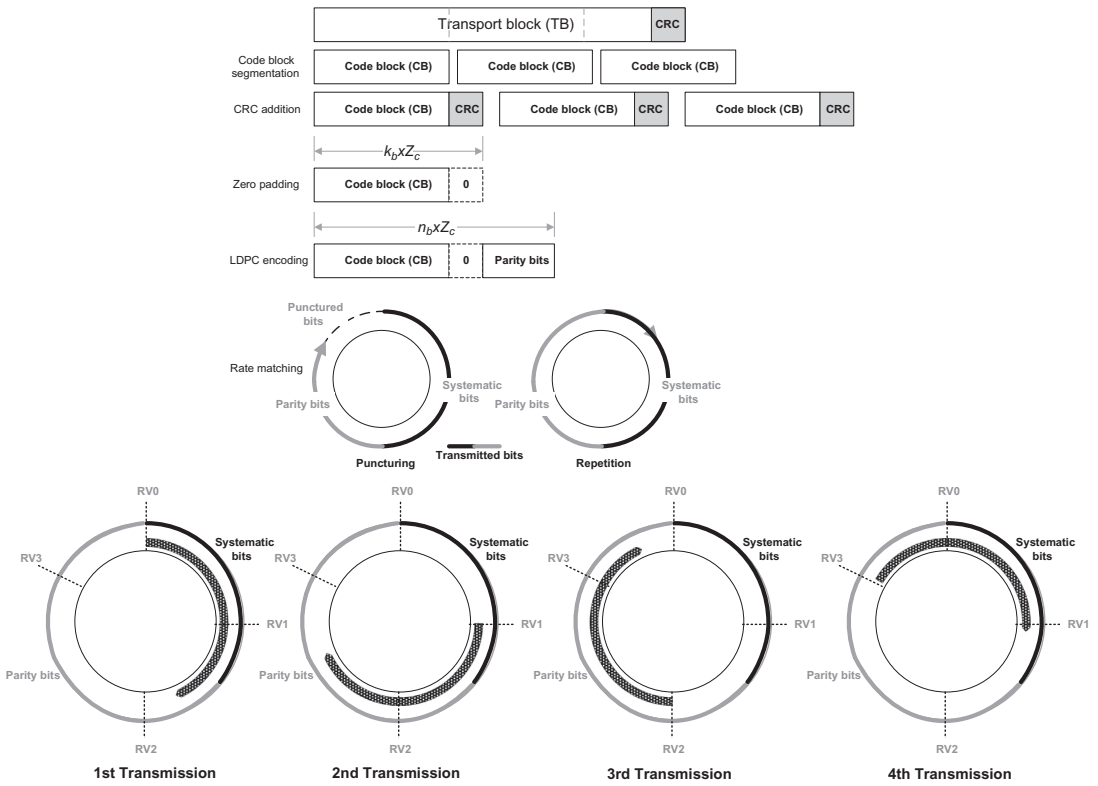


Figure 4.36
Example of rate-matching and code block concatenation processes [14].

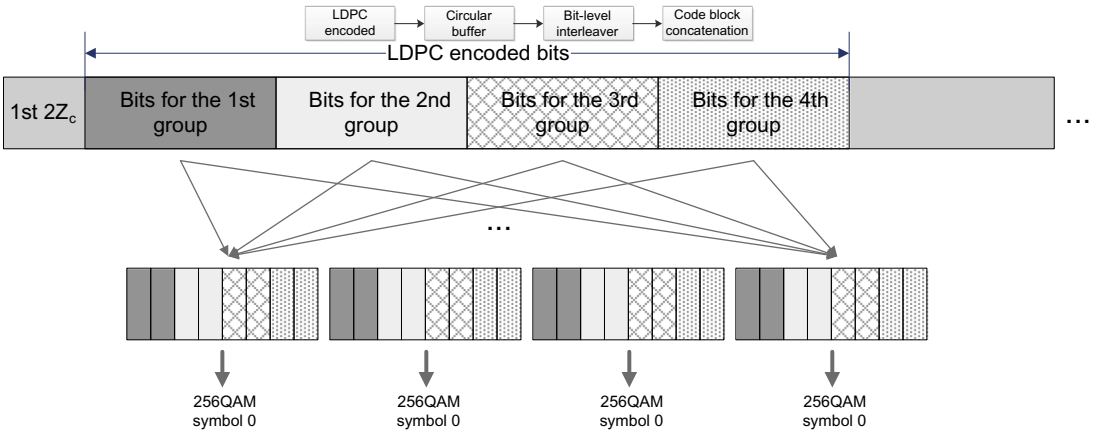


Figure 4.37
Example bit-level interleaving for 256QAM modulation.

Table 4.9: Starting position of different redundancy versions [7].

rv_{id}	k_0	
	LDPC Base Graph 1	**LDPC Base Graph 2**
0	0	0
1	$\left\lfloor \frac{17N_{cb}}{66Z_c} \right\rfloor Z_c$	$\left\lfloor \frac{13N_{cb}}{50Z_c} \right\rfloor Z_c$
2	$\left\lfloor \frac{33N_{cb}}{66Z_c} \right\rfloor Z_c$	$\left\lfloor \frac{25N_{cb}}{50Z_c} \right\rfloor Z_c$
3	$\left\lfloor \frac{56N_{cb}}{66Z_c} \right\rfloor Z_c$	$\left\lfloor \frac{43N_{cb}}{50Z_c} \right\rfloor Z_c$

required number of bits from the circular buffer where the exact set of bits to transmit depends on the RV corresponding to different starting positions in the circular buffer. Thus by selecting different RVs, different sets of coded bits representing the same set of information bits can be generated, which is used when implementing HARQ with incremental redundancy. The starting points in the circular buffer (RV0,RV1,RV2,RV3) are defined such that both RV0 and RV3 codes are self-decodable which means that they include the systematic bits under typical conditions. The RV index of the incremental redundancy HARQ in NR is derived differently compared to LTE. Unlike LTE that RV index positions are sequentially incremented in the circular buffer, in NR, if $rv_{id} = 0, 1, 2, 3$ denotes the RV number for the current transmission, the rate matching output bit sequence $\{e_k | k = 0, 1, \ldots, E-1\}$ is generated as $e_k = d_{(k_0+j) \bmod N_{cb}}$ where k_0 is given by Table 4.9 according to the value of rv_{id} and the LDPC base graph lifting factor [7,14].

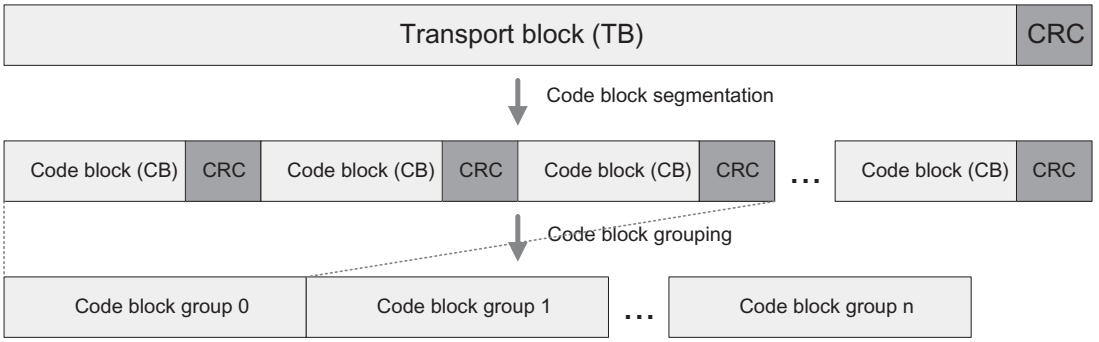


Figure 4.38
CBG-based retransmissions [32].

It is possible to perform HARQ retransmissions with a code block granularity. In that case the information included in the DCI would determine the code block group (CBG) which is (re)transmitted, and information about handling the CBGs for soft-buffer/HARQ combining purposes (see Fig. 4.38). The NR supports code block group-based transmission with single or multi-bit HARQ-ACK feedback. For the case of CBG-based retransmission, HARQ-ACK multiplexing is supported. The motivation for CBG-based retransmission is to improve the spectrum efficiency because if CBG-based retransmission is configured, the HARQ feedback is provided per CBG and only the erroneously received code block groups are retransmitted. This can consume less radio resources than retransmitting the entire TB. If the retransmission is caused due to low SNR then combining in the soft-buffer would improve the decoding quality during retransmissions; however, if the retransmitted code block was affected by preemption, the buffer content is not correct, and it is better to discard the content of the buffer and to request a fresh transmission.

If a UE is configured to receive CBG-based transmissions when the RRC parameter *codeBlockGroupTransmission* set for PDSCH, it determines the number of CBGs for a PDSCH transmission by calculating $M = \min(N_{CBG}, C)$ where N_{CBG} denotes the maximum number of CBGs per TB which is configured by RRC parameter *maxCodeBlockGroups PerTransportBlock* for PDSCH, and C is the number of code blocks. We define $M_1 = C \bmod M, K_1 = \lceil C/M \rceil$ and $K_2 = \lfloor C/M \rfloor$. If $M_1 > 0$, the mth CBG when $m = 0, 1, \ldots, M_1 - 1$ consists of code blocks with indices $mK_1 + k, \forall k = 0, 1, \ldots, K_1 - 1$. The mth CBG when $m = M_1, M_1 + 1, \ldots, M - 1$ consists of code blocks with indices $M_1K_1 + (m - M_1)K_2 + k, \forall k = 0, 1, \ldots, K_2 - 1$. If a UE is configured with CBG-based retransmissions, the scheduling assignment for the UE would contain the necessary HARQ-related control signaling including process number, new-data indicator, CBG transmit indicator (CBGTI), and the CBG flush indicator (CBGFI) as well as information to handle the transmission of the HARQ acknowledgment in the uplink such as timing and

resource indication information. Upon receiving a scheduling assignment in the DCI, the receiver would attempt to decode the TB. Since transmissions and retransmissions are scheduled using the same framework, the UE needs to know whether this is a new transmission, in that case the soft buffer should be flushed, or a retransmission, where soft combining should be performed. Therefore, a single-bit new data indicator is included as part of the scheduling information. The new data indicator operates at TB level. However, if CBG-based retransmissions are configured, the device needs to know which CBGs are retransmitted and whether the corresponding soft buffer should be flushed. This is handled by additional information fields in the DCI when CBG-based retransmissions are configured, that is, CBGTI and CBGFI fields. The CBGTI is a bitmap indicating whether a certain CBG is present in the downlink transmission. The CBGFI is a single bit field, indicating whether the CBGs identified by CBGTI should be flushed or whether soft combining should be performed. The decoding operation results in either a positive acknowledgment in the case of a successful decoding or a negative acknowledgment in the case of unsuccessful decoding, and it is fed back to the gNB as part of the uplink control information. If CBG-based retransmissions are configured, a bitmap with one bit per CBG is fed back instead of a single bit representing the entire TB. The uplink uses the same asynchronous HARQ protocol as the downlink. The necessary HARQ-related information including process number, new-data indicator, and CBGTI (when configured) are included in the scheduling grant [9]. CBG-based retransmissions are transparent to the MAC sublayer and are handled in the physical layer despite being part of the HARQ mechanism. From the MAC perspective, the TB is not correctly received until all CBGs are correctly received and decoded. It is not possible to combine transmission of new CBGs associated with another TB with retransmissions of CBGs belonging to the incorrectly received TB in the same HARQ process [14].

The bit sequence $f_0, f_1, \ldots, f_{E-1}$ is generated by interleaving bit sequence $e_0, e_1, \ldots, e_{E-1}$ according to the value of the modulation order Q_m as follows $f_{i+jQ_m} = e_{iE/Q_m+j}$, $\forall j = 0, 1, \ldots, E/Q_m - 1; i = 0, 1, \ldots, Q_m - 1$ [7]. The new radio supports up to two codewords in the downlink transmission. For each codeword cw, the block of bits $b^{(cw)}(0), b^{(cw)}(1), \ldots, b^{(cw)}(N_{bit}^{(cw)} - 1)$, where $N_{bit}^{(cw)}$ denotes the number of bits in codeword cw transmitted on the physical shared channel, is scrambled prior to modulation, resulting in a block of scrambled bits $\tilde{b}^{(cw)}(0), \tilde{b}^{(cw)}(1), \ldots, \tilde{b}^{(cw)}(N_{bit}^{(cw)} - 1)$ such that $\tilde{b}^{(cw)}(i) = [b^{(cw)}(i) + c^{(cw)}(i)] \bmod 2$. The scrambling sequence $c^{(cw)}(i)$ is a generic pseudo-random length-31 Gold sequence that is initialized by setting $c_{init} = n_{RNTI}2^{15} + cw2^{14} + n_{ID}$ where $n_{ID} \in \{0, 1, \ldots, 1023\}$, if configured through RRC signaling; otherwise $n_{ID} = N_{ID}^{cell}$, n_{RNTI} corresponds to the RNTI associated with the PDSCH transmission [6].

The scrambled bit sequence $\tilde{b}^{(cw)}(0), \tilde{b}^{(cw)}(1), \ldots, \tilde{b}^{(cw)}(N_{bit}^{(cw)} - 1)$ is modulated using one of the modulation schemes, for example, QPSK, 16QAM, 64QAM, or 256QAM. The complex-valued modulation symbols for each of the codewords that are going to be transmitted are mapped to one or several layers for spatial multiplexing. As shown in Table 4.10, the

Table 4.10: Codeword to layer mapping for spatial multiplexing [6].

Number of Layers v	Number of Codewords cw	Mapping	Parameters
1	1	$x^{(0)}(i) = d^{(0)}(i)$	$N_{layer}^{symb} = N_{symb}^{(0)}; i = 0, 1, \ldots, N_{layer}^{symb} - 1$
2	1	$x^{(0)}(i) = d^{(0)}(2i)$ $x^{(1)}(i) = d^{(0)}(2i+1)$	$N_{layer}^{symb} = N_{symb}^{(0)}/2$
3	1	$x^{(0)}(i) = d^{(0)}(3i)$ $x^{(1)}(i) = d^{(0)}(3i+1)$ $x^{(2)}(i) = d^{(0)}(3i+2)$	$N_{layer}^{symb} = N_{symb}^{(0)}/3$
4	1	$x^{(0)}(i) = d^{(0)}(4i)$ $x^{(1)}(i) = d^{(0)}(4i+1)$ $x^{(2)}(i) = d^{(0)}(4i+2)$ $x^{(3)}(i) = d^{(0)}(4i+3)$	$N_{layer}^{symb} = N_{symb}^{(0)}/4$
5	2	$x^{(0)}(i) = d^{(0)}(2i)$ $x^{(1)}(i) = d^{(0)}(2i+1)$ $x^{(2)}(i) = d^{(1)}(3i)$ $x^{(3)}(i) = d^{(1)}(3i+1)$ $x^{(4)}(i) = d^{(1)}(3i+2)$	$N_{layer}^{symb} = N_{symb}^{(0)}/2 = N_{symb}^{(1)}/3$
6	2	$x^{(0)}(i) = d^{(0)}(3i)$ $x^{(1)}(i) = d^{(0)}(3i+1)$ $x^{(2)}(i) = d^{(0)}(3i+2)$ $x^{(3)}(i) = d^{(1)}(3i)$ $x^{(4)}(i) = d^{(1)}(3i+1)$ $x^{(5)}(i) = d^{(1)}(3i+2)$	$N_{layer}^{symb} = N_{symb}^{(0)}/3 = N_{symb}^{(1)}/3$
7	2	$x^{(0)}(i) = d^{(0)}(3i)$ $x^{(1)}(i) = d^{(0)}(3i+1)$ $x^{(2)}(i) = d^{(0)}(3i+2)$ $x^{(3)}(i) = d^{(1)}(4i)$ $x^{(4)}(i) = d^{(1)}(4i+1)$ $x^{(5)}(i) = d^{(1)}(4i+2)$ $x^{(6)}(i) = d^{(1)}(4i+3)$	$N_{layer}^{symb} = N_{symb}^{(0)}/3 = N_{symb}^{(1)}/4$
8	2	$x^{(0)}(i) = d^{(0)}(4i)$ $x^{(1)}(i) = d^{(0)}(4i+1)$ $x^{(2)}(i) = d^{(0)}(4i+2)$ $x^{(3)}(i) = d^{(0)}(4i+3)$ $x^{(4)}(i) = d^{(1)}(4i)$ $x^{(5)}(i) = d^{(1)}(4i+1)$ $x^{(6)}(i) = d^{(1)}(4i+2)$ $x^{(7)}(i) = d^{(1)}(4i+3)$	$N_{layer}^{symb} = N_{symb}^{(0)}/4 = N_{symb}^{(1)}/4$

complex-valued modulation symbols $d^{(cw)}(0), d^{(cw)}(1), \ldots, d^{(cw)}(N_{symb}^{(cw)} - 1)$ corresponding to codeword cw are mapped to layers $x(i) = \left[x^{(0)}(i) \quad \ldots \quad x^{(v-1)}(i) \right]^T$, $\forall i = 0, 1, \ldots, N_{layer}^{symb} - 1$ where v denotes the number of layers and N_{layer}^{symb} is the number of modulation symbols per layer [6].

The block of vectors $x(i) = [x^{(0)}(i) \ldots x^{(v-1)}(i)]^T, \forall i = 0, 1, \ldots, N_{layer}^{symb} - 1$ are mapped to antenna ports as follows:

$$[y^{(p_0)}(i) \ldots y^{(p_{v-1})}(i)] = [x^{(0)}(i) \ldots x^{(v-1)}(i)], \forall i = 0, 1, \ldots, N_{layer}^{symb} - 1$$

The set of antenna ports $\{p_0, p_1, \ldots, p_{v-1}\}$ are determined according to the procedure specified in [7].

For each antenna port that is used for transmission of the physical shared channel, the properly scaled block of complex-valued symbols $y^{(p)}(0), y^{(p)}(1), \ldots, y^{(p)}(N_{AP}^{symb} - 1)$ are sequentially mapped to the virtual resource elements (k', l) allocated for transmission of PDSCH that have not been designated for reference signals. Any common resource block partially or fully overlapping with an SS/PBCH block is considered occupied and is not used for transmission. The virtual resource elements are mapped in frequency-first manner as illustrated in Fig. 4.39. The virtual resource blocks are mapped to physical resource blocks in

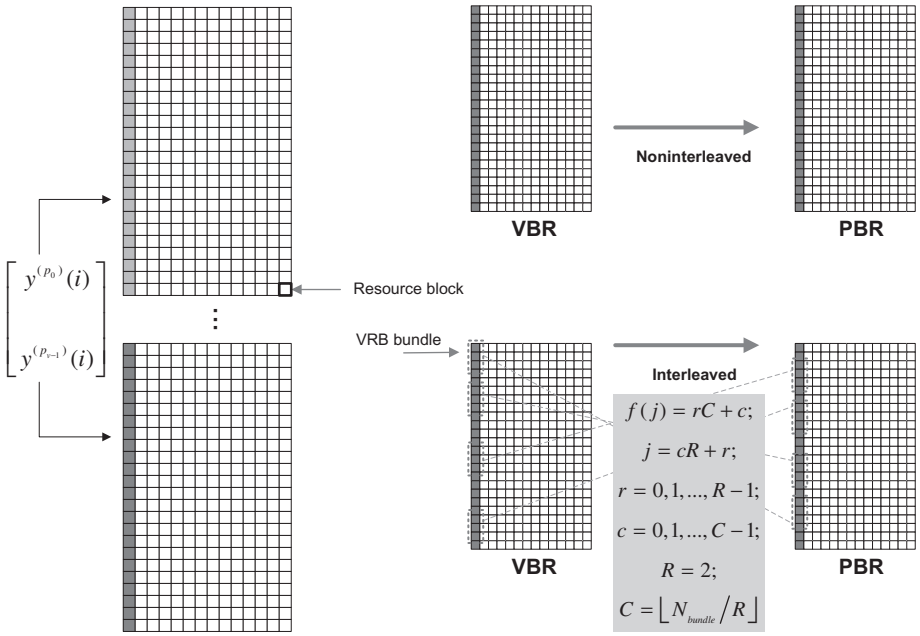


Figure 4.39
Mapping to VRBs and from VRBs to PRBs [6].

form of either non-interleaved or interleaved, wherein the non-interleaved is the default mapping scheme.

For non-interleaved VRB-to-PRB mapping, the virtual resource block n is mapped to physical resource block n, except for PDSCH transmissions scheduled with DCI format 1_0 in a common search space where virtual resource block n is mapped to physical resource block $n + N_{CORESET}^{start}$, where $N_{CORESET}^{start}$ is the lowest numbered physical resource block in the CORESET where the corresponding DCI was received [6].

In interleaved mapping scheme, the mapping process is defined in terms of resource block bundles. The set of $N_{BWP}^{size}(i)$ resource blocks in the ith bandwidth part with starting position $N_{BWP}^{start}(i)$ are divided into $N_{bundle} = \left\lceil [N_{BWP}^{size}(i) + (N_{BWP}^{start}(i) \bmod L_i)]/L_i \right\rceil$ resource-block bundles in increasing order of the resource-block number and bundle number where L_i is the bundle size for the ith bandwidth part defined by RRC parameter *vrb-ToPRB-Interleaver* and resource block bundle 0 consists of $L_i - \left(N_{BWP}^{start}(i) \bmod L_i\right)$ resource blocks, resource block bundle $N_{bundle} - 1$ consists of $(N_{BWP}^{start}(i) + (N_{BWP}^{size}(i)) \bmod L_i, \forall L_i > 0$ resource blocks; otherwise, all resource block bundles consists of L_i resource blocks (except for PDSCH transmissions scheduled with DCI format 1_0 with the CRC scrambled by SI-RNTI in Type0-PDCCH common search space in CORESET 0 and in any common search space other than Type0-PDCCH common search space). The virtual resource blocks in the region $j \in \{0, 1, \ldots, N_{bundle} - 2\}$ are mapped to the physical resource blocks such that virtual resource block bundle $N_{bundle} - 1$ is mapped to physical resource block bundle $N_{bundle} - 1$ and virtual resource block bundle $j \in \{0, 1, \ldots, N_{bundle} - 2\}$ is mapped to physical resource block bundle $f(j)$ where $f(j) = rC + c; j = cR + r; r = 0, 1, \ldots, R - 1; c = 0, 1, \ldots, C - 1; R = 2;$ and $C = \left\lceil N_{bundle}/R \right\rceil$. If no bundle size is configured, the UE will assume $L_i = 2$ with a precoding resource block group (PRG) size of 4 (see Fig. 4.39).

4.1.6 CSI Measurement and Reporting and Beam Management

4.1.6.1 CSI Measurement and Reporting

In wireless communications, the channel state information refers to channel properties of a wireless communication link. This information describes how a signal propagates from the transmitter to the receiver and represents the combined effect of scattering, multipath fading, signal power attenuation with distance, etc. The knowledge of CSI at the transmitter and/or the receiver makes it possible to adapt data transmission to current channel conditions, which is crucial for achieving reliable and robust communication with high data rates in multi-antenna systems. The CSI is often required to be estimated at the receiver, and usually quantized and fed back to the transmitter. The downlink channel can be estimated from uplink reference signals in TDD systems under certain conditions due to reciprocity.

In general, the transmitter and receiver can observe different CSI. There are two types of CSI, that is, instantaneous CSI and statistical CSI. In instantaneous CSI or short-term CSI the current channel conditions are known, which can be interpreted as knowing the impulse response of a digital filter. This provides an opportunity to adapt the transmit signal to the impulse response and thereby to optimize the received signal for spatial multiplexing or to achieve low bit-error-rates. In statistical CSI or long-term CSI, the statistical characteristics or statistics of the channel are known. The latter information may include the type of fading distribution, the average channel gain, the line-of-sight component, and the spatial correlation. Similar to the instantaneous CSI, this information can be used for optimization of transmission parameters. The CSI estimation accuracy is practically limited by how fast the channel conditions are varying. In fast-fading channels where the channel conditions may vary rapidly during transmission of a single information symbol, only statistical CSI is reasonable. On the other hand, in slow-fading scenarios, the instantaneous CSI can be estimated with reasonable precision and used for transmission adaptation for a period of time before becoming obsolete. In practical scenarios, the available CSI is often manifested as instantaneous CSI with some estimation/quantization error combined with some statistical information.

To support diverse use cases, NR features a highly flexible and unified CSI framework, in which there is reduced coupling between CSI measurements, CSI reporting and the actual downlink transmission compared to LTE. The CSI framework can be seen as a toolbox, where different CSI reporting settings and CSI-RS resource settings for channel and interference measurements can be selected, so that they correspond to the antenna configuration and transmission scheme in use such that the CSI reports on different beams can be dynamically triggered. The framework also supports more advanced schemes such as multi-point transmission and coordination. The control and data transmissions follow a self-contained principle, where all information required to decode the transmission (such as accompanying DM-RS) is contained within the transmission itself. As a result, the network can seamlessly change the transmission point or beam as the UE moves in the network. The CSI-RS reference signals are used for CSI acquisition and beam management. The CSI-RS resources for a UE are configured by RRC information elements and can be dynamically activated/deactivated via MAC control elements or DCI [57].

The configuration and use of CSI-RS in NR can be defined via the CSI framework. As shown in Fig. 4.40, the basic units of CSI framework in NR are CSI reporting setting and CSI resource setting. The CSI reporting setting is linked to M resource settings for channel and interference measurements (CM and IM), where $M = 1$ indicates resource setting for channel measurement and beam management; $M = 2$ indicates resource settings for channel measurements and CSI-interference measurement (CSI-IM) or NZP CSI-RS for interference measurement; and $M = 3$ is an indication for resource settings for channel measurements and two resource settings for CSI-IM and NZP CSI-RS−based interference measurement.

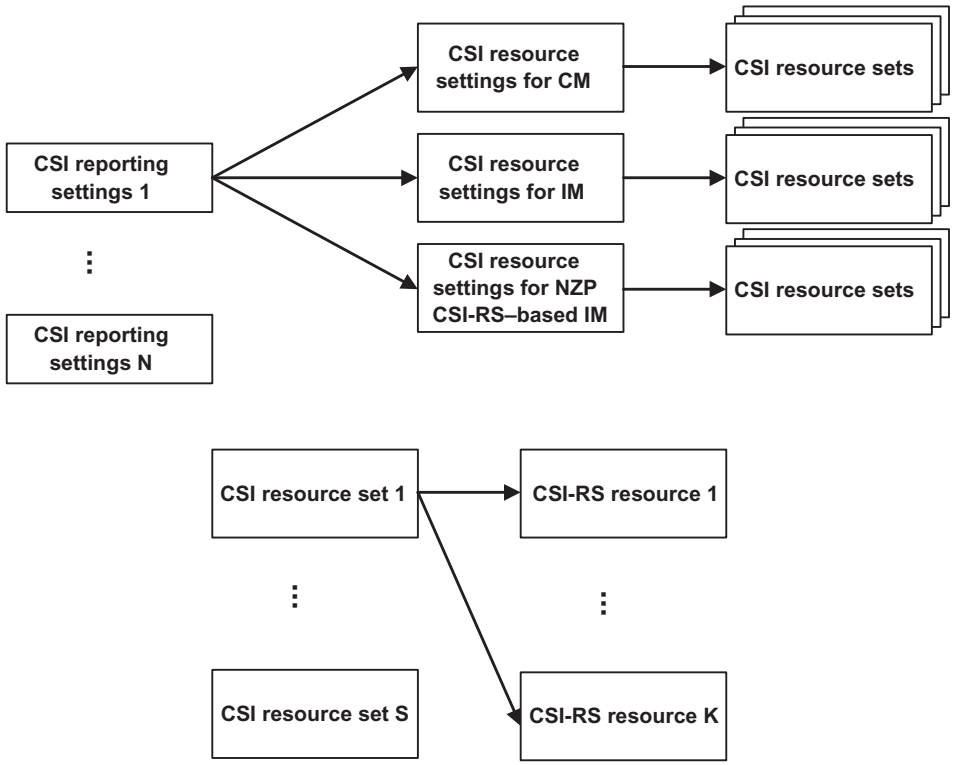


Figure 4.40
CSI framework in NR.

The above-mentioned resource settings are linked to *S* resource sets each resource set comprises SS/PBCH block resources for beam management and is linked to CSI-RS resources [9].

The time and frequency resources that can be used by the UE to report CSI are controlled by the gNB. The CSI may consist of CQI, PMI, CSI-RS resource indicator (CRI), SS block resource indicator, layer indication (LI), rank indicator (RI), and/or and L1-RSRP measurements. For CQI, PMI, CRI, LI, RI, L1-RSRP, the UE is configured via RRC signaling with more than one *CSI-ReportConfig* reporting settings, *CSI-ResourceConfig* resource settings, and one or two lists of trigger states, indicating the resource set IDs for channel and optionally for interference measurement. Each trigger state contains an associated *CSI-ReportConfig* [9].

Each reporting setting *CSI-ReportConfig* is associated with a single downlink BWP and contains the reported parameter(s) for one CSI reporting band including CSI Type-I or II, codebook configuration comprising codebook subset restriction, time-domain behavior,

frequency granularity for CQI and PMI, measurement restriction configurations, LI, reported L1-RSRP parameter(s), CRI, and the SSB resource indicator. The time-domain behavior of the *CSI-ReportConfig* is determined by RRC signaling and can be set to aperiodic, semi-persistent, or periodic. For periodic and semi-persistent CSI reporting, the configured periodicity and slot offset applies in the numerology of the uplink BWP in which the CSI report is configured to be transmitted. The higher layer parameter *ReportQuantity* identifies the CSI-related or L1-RSRP-related quantities to report. Another RRC parameter indicates the reporting granularity in the frequency domain including the CSI reporting band and whether PMI/CQI reporting is wideband or subband. The *CSI-ReportConfig* can also contain *CodebookConfig*, which contains configuration parameters for Type-I or Type-II CSI including codebook subset restriction, and configurations of group-based reporting [9].

Each CSI resource setting contains a configuration of more than one CSI resource sets, each consisting of CSI-RS resources (either NZP CSI-RS or CSI-IM) and SS/PBCH block resources used for L1-RSRP computation. Each CSI resource setting located in the downlink BWP is defined by RRC signaling, and all CSI resource settings are linked to a CSI report setting within the same downlink BWP. The reporting configuration for CSI can be aperiodic (using PUSCH), periodic (using PUCCH), or semi-persistent (using PUCCH and DCI activated PUSCH). The CSI-RS resources can be periodic, semipersistent, or aperiodic. The supported combinations of CSI reporting configurations and CSI-RS resource in NR are shown in Tables 4.11 and 4.12. If interference measurement is performed using CSI-IM, each CSI-RS resource for CM is resource-wise associated with a CSI-IM resource based on the ordering of the CSI-RS resource and CSI-IM resource in the corresponding resource sets. The number of CSI-RS resources for channel measurement equals to the number of CSI-IM resources [9].

The CSI reports are used to provide the gNB with an estimate of the downlink communication channel observed by the UE in order to assist channel-dependent scheduling. The new

Table 4.11: Triggering/activation of CSI reporting for CSI-RS configurations [9].

CSI-RS Configuration	Periodic CSI Reporting	Semi-persistent CSI Reporting	Aperiodic CSI Reporting
Periodic CSI-RS	No dynamic triggering/ activation	For reporting on PUCCH, the UE receives an activation command and for reporting on PUSCH, the UE receives triggering on DCI	Triggered by DCI or activation command
Semi-persistent CSI-RS	Not supported	For reporting on PUCCH, the UE receives an activation command, whereas for reporting on PUSCH, the UE receives the trigger on DCI	Triggered by DCI or activation command
Aperiodic CSI-RS	Not supported	Not supported	Triggered by DCI or activation command

Table 4.12: Major components of NR CSI framework [9].

CSI Report Settings	CSI Resource Settings	CSI Trigger States
It defines what CSI to report and when to report it. • Quantities to report: CSI related or L1-RSRP related • Time-domain behavior: aperiodic, semi-persistent, periodic • Frequency-domain granularity: reporting band, wideband, subband • Time-domain restrictions: For channel and interference measurements • Codebook configuration parameters: Type-I and Type-II	It defines what signals to use to compute CSI. • A resource setting configures more than one CSI resource sets where each CSI resource set consists of CSI-RS resources (either NZP CSI-RS or CSI-IM); and SS/PBCH Block Resources that are used for L1-RSRP calculation • Time-domain behavior: aperiodic, semi-persistent, periodic as well as periodicity and slot offset *Note: The number of CSI-RS Resource Sets is limited to one, if CSI Resource Setting is periodic or semi-persistent*	It associates "what CSI to report and when to report it" with "what signals to use to compute CSI." • Links report settings with resource settings • Contains the list of associated *CSI-ReportConfig*

radio supports analog beamforming and high-resolution CSI feedback through beam management, where a UE measures a set of analog beams for each digital port and reports the beam quality. The gNB then assigns a number of analog beams to the UE. As the downlink channel experienced by the UE varies, the gNB can change this assignment when the link associated with an assigned beam deteriorates. While beam management is especially instrumental in above 6 GHz frequency bands, it can also be applied to sub-6 GHz bands. Furthermore, NR supports a modular and scalable CSI framework, where high-resolution spatial channel information is provided via two-stage precoding. The first stage involves the choice of a basis subset, and the second stage incorporates a set of coefficients for approximating the channel eigenvector with a linear combination of the basis subsets. It must be noted that while beam management and CSI acquisition can be independently operated, they can be used together to support mobile UEs.

For UEs in RRC_CONNECTED state, in addition to the SS block, UE-specific CSI-RS can be configured in order to improve the quality of UE measurements and to provide better user-centric mobility experience. For example, in high-frequency bands, narrow-beam CSI-RS can be configured for the UEs at the edge of the cell in order to achieve better signal-to-interference-plus-noise ratio (SINR) range and measurement accuracy. As shown in Fig. 4.41,

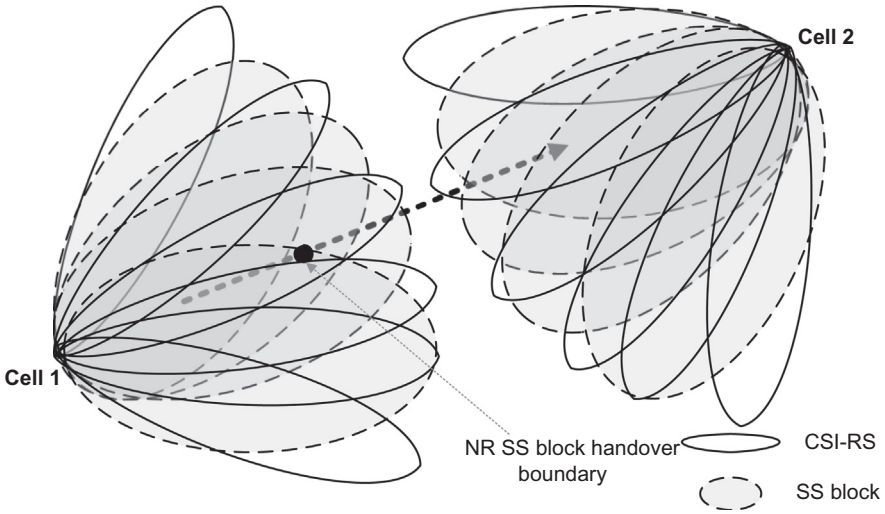


Figure 4.41
Configurable CSI-RS for downlink mobility measurements [72].

assuming the same energy per resource element (EPRE) is applied to CSI-RS and SSB resources for transmission, narrow-beam CSI-RS measurement can provide better SINR range which can improve RSRP measurement accuracy compared to wide-beam SSB measurements. The CSI-RS properties of the serving and neighboring cells for the mobility measurements can include NR cell ID, slot configuration used to obtain the slot offset for CSI-RS and the periodicity, for example, 5, 10, 20, 40 ms, configurable measurement bandwidth of CSI-RS, configurable parameter for CSI-RS scrambling sequence, configurable numerologies, and association between CSI-RS for mobility measurement and SSB, such as spatial QCL[8] information. The above CSI-RS properties are signaled to the UE via dedicated RRC signaling [72].

[8] Two antenna ports are said to be quasi co-located, if the properties of the channel over which a symbol on one antenna port is transmitted can be inferred from the channel over which a symbol on the other antenna port is transmitted. The QCL supports beam management (spatial parameter), frequency/timing offset estimation (Doppler/delay), and RRM measurements (average gain). The reference signal set contains a reference to either one or two downlink reference signals and an associated quasi co-location type (QCL-Type) for each one configured by an RRC parameter. The quasi co-location relationship is configured by the RRC parameter *qcl-Type1* for the first downlink reference signal, and *qcl-Type2* for the second downlink reference signal (if configured). The quasi co-location types corresponding to each reference signal are given by the RRC parameter *qcl-Type* in *QCL-Info* and may take one of the following values [9]:

- QCL-TypeA': {Doppler shift, Doppler spread, average delay, delay spread}
- QCL-TypeB': {Doppler shift, Doppler spread}
- QCL-TypeC': {average delay, Doppler shift}
- QCL-TypeD': {Spatial RX parameter}

The NR supports two types of spatial-resolution CSI: standard-resolution (Type I) and high--resolution (Type II). The low-resolution CSI is targeted for SU-MIMO transmission since it relies on the UE receiver to suppress the inter-layer interference. This is possible since the number of received layers is less than the number of receiver antennas for a given UE. For MU-MIMO transmission, the number of received layers is typically larger than the number of receive antennas for the UE. The base station exploits beamforming/precoding to suppress inter-UE interference. Thus a higher resolution CSI, capturing more propagation paths of the channel, is needed to provide sufficient degrees of freedom at the transmitter [57].

In LTE, the UEs are configured with a transmission mode and a number of CSI reporting modes which are limited by complexity and scalability, whereas in NR, a modular framework is specified where a UE can be configured with one measurement setting, which includes $N \geq 1$ CSI reporting and $M \geq 1$ CSI resource settings. A CSI resource setting can be associated with one or more reporting settings to flexibly support beam management and CSI acquisition, resulting in $L \geq 1$ links. A UE can be dynamically assigned one or more reporting settings or links to generate the desired CSI report, which may include CRI, which is used to indicate a preferred CSI-RS resource from a configured set since different CSI-RS resources in the set can be differently precoded, rank indicator (RI), CQI, and PMI. The CRI, RI, CQI, and PMI are associated with resource selection (when a UE measures multiple CSI-RS resources), the number of dominant downlink channel directions, sustained spectral efficiency or related SINR values, and the dominant channel directions chosen from a codebook of vectors or matrices. Since CSI requirements for different operational modes are different, the CSI reporting settings can include different CSI components for CSI acquisition (see Fig. 4.40 and Table 4.12).

The UE measures the spatial channel between itself and the serving base station using the CSI-RS transmitted from the gNB transmit antenna ports in order to generate a CSI report. The UE then calculates the CSI-related metrics and reports the CSI to the gNB. Using the reported CSIs from all UEs, the gNB performs link adaptation and scheduling. The goal of CSI measurement and reporting is to obtain an approximation of the CSI. This can be achieved when the reported PMI accurately represents the dominant channel eigenvector(s), thereby enabling accurate beamforming.

The standard-resolution (Type I) CSI utilizes a dual-stage codebook with precoding matrix $\mathbf{W} = \mathbf{W_1}\mathbf{W_2}$ incorporating a wideband $\mathbf{W_1}$ matrix that is common for all subbands, capturing long-term channel characteristics, and a subband $\mathbf{W_2}$ matrix representing fast fading properties of the channel. In this context a subband comprises multiple consecutive resource blocks. Type I codebooks support up to rank 8, that is, the rank indicator $RI \in \{1, 2, \ldots, 8\}$. Designed for N_{panel} panels of dual-polarized arrays, the $\mathbf{W_1}$ matrix factor is constructed from $2N_{panel}$ blocks of two-dimensional DFT matrices. The $\mathbf{W_2}$ matrix selects a subset of DFT vectors from $\mathbf{W_1}$ and applies phase shifts (taken from the phase shift keying alphabet)

across panels and polarizations. For Type I, when $RI = 1$, the selected subset includes only one DFT vector, and when $RI > 1$, the subset includes multiple DFT vectors to generate orthogonal DFT beams. The phase shifts are introduced across two polarizations since the associated channels tend to be uncorrelated. This is also utilized for multi-panel arrays since the spacing between the last element of a panel and the first element of the next panel is different from the inter-element spacing within a panel. In addition, the panels may not be sufficiently phase- and/or timing-calibrated [57].

For high-spatial-resolution (Type II) CSI, feedback of up to two layers, that is, $RI \in \{1, 2\}$ is supported with a linear combination codebook. The codebook resolution is sufficiently high to facilitate sufficiently accurate approximation of the downlink channel. In this scheme the UE reports a PMI that represents a linear combination of multiple beams, as shown in Fig. 4.42. Similar to Type I, Type II employs a dual-stage $\mathbf{W} = \mathbf{W_1}\mathbf{W_2}$ codebook wherein $\mathbf{W_1}$ is a wideband and $\mathbf{W_2}$ is a subband precoder. However, unlike Type I, the recommended precoding matrix from Type II CSI is non-constant modulus since different precoder elements can have different magnitudes [57].

For Type II CSI and $RI \in \{1, 2\}$, the $\mathbf{W_1}$ matrix selects a subset of DFT vectors of size $L \in \{2, 3, 4\}$ which serves as a basis set for linear combination performed by $\mathbf{W_2}$. This subset selection is common across two polarizations and, for $RI = 2$, two transmission layers. The linear combination is performed per subband as well as independently across polarizations and layers. To reduce the feedback overhead for $\mathbf{W_2}$, some partial information pertaining to linear combination such as the strongest of $2L$ linear combination coefficients and $2L - 1$ wideband reference amplitudes for subband differential encoding of the linear combination coefficients in $\mathbf{W_2}$ is also included in $\mathbf{W_1}$. Therefore, the amplitude component of the linear combination coefficients comprises wideband and subband components. The phase component is per subband and configurable as QPSK or 8-PSK. Due to large degrees of freedom offered by Type II CSI, the number of precoder hypotheses is large. However, exhaustive codebook search, which is prohibitively complex, is not needed. Due to high spatial resolution, the precoder can be efficiently determined by performing scalar quantization of each of the channel eigenvector coefficients. Since Type II CSI is configurable in terms of its basis set size $L \in \{2, 3, 4\}$, the amplitude frequency granularity, that is, wideband-only or wideband + subband, and phase shift (QPSK or 8-PSK), a range of performance-overhead trade-offs would be possible [57].

There are two subtypes of Type I CSI that are referred to as Type I single-panel CSI and Type I multi-panel CSI, corresponding to different codebooks. These codebooks are designed assuming different antenna configurations on the gNB transmitter. The codebooks for Type I single-panel CSI are designed assuming a single antenna panel with $N_H \times N_V$ cross-polarized antenna elements. In general, the precoder matrix $\mathbf{W}$ for Type I single-panel CSI can be constructed as the product of two matrices $\mathbf{W_1}$ and $\mathbf{W_2}$ where the information

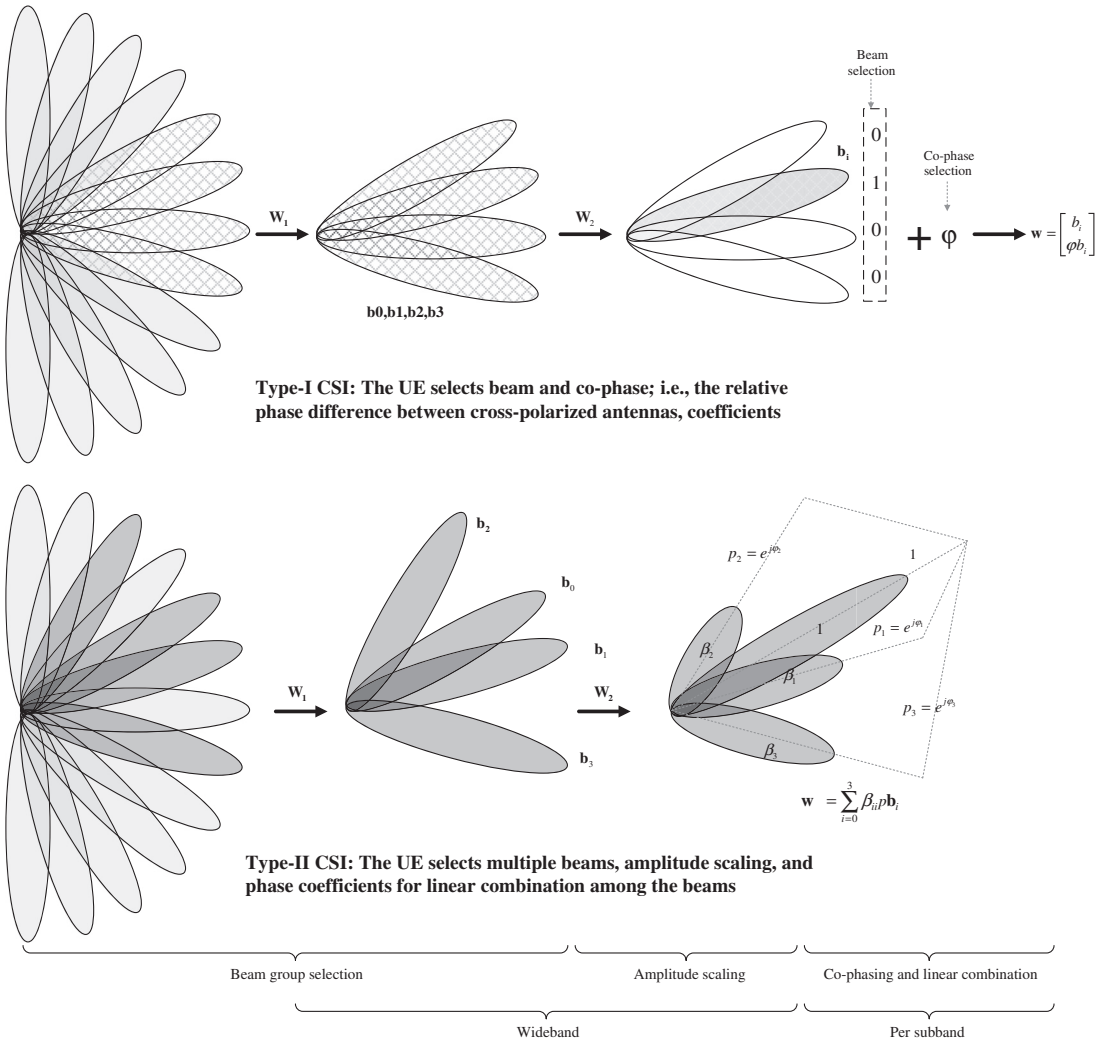


Figure 4.42
Illustration of Type-I and Type-II CSI in NR [57,69].

about the selected $\mathbf{W_1}$ and $\mathbf{W_2}$ is reported separately in different parts of the PMI. The matrix $\mathbf{W_1}$ captures long-term frequency-independent characteristic of the channel. A single $\mathbf{W_1}$ is selected and reported for the entire reporting bandwidth (wideband feedback). In contrast, the matrix $\mathbf{W_2}$ encompasses short-term frequency-dependent characteristic of the channel. Thus the precoder matrix can be selected and reported on a subband basis where a subband covers a fraction of the overall reporting bandwidth. Alternatively, the device may decide not to report $\mathbf{W_2}$ when subsequently selecting CQI. In that case, it should assume that the network randomly selects $\mathbf{W_2}$ on a per physical resource block group basis.

Note that this does not impose any restrictions on the actual precoding applied at the gNB side, rather it is only about the assumptions that the device would make when selecting CQI. The matrix $\mathbf{W_1}$ can be considered as defining a beam or a group of beams pointing toward a specific direction. More specifically, the matrix $\mathbf{W_1}$ can be written as $\mathbf{W_1} = \begin{bmatrix} \mathbf{b} & \mathbf{0} \\ \mathbf{0} & \mathbf{b} \end{bmatrix}$ where each column of the matrix $\mathbf{b}$ defines a beam. The 2×2 structure of the matrix corresponds to two polarizations. Note that, as the matrix $\mathbf{W_1}$ is assumed to represent long-term frequency-independent channel characteristics, the same beam direction can be assumed to fit both polarization directions. Selecting matrix $\mathbf{W_1}$ or equivalently $\mathbf{b}$ can be seen as selecting a specific beam direction from a large set of possible beam directions defined by the full set of $\mathbf{W_1}$ matrices within the codebook. In the case of rank 1 or rank 2 transmissions, either a single beam or four adjacent beams are defined by the matrix $\mathbf{W_1}$ (see Fig. 4.42). In the case of four adjacent beams corresponding to four columns in matrix $\mathbf{b}$, matrix $\mathbf{W_2}$ would select the exact beam to be used for the transmission. Since $\mathbf{W_2}$ can be reported on a subband basis, it is possible to adjust the beam direction per subband. In addition, $\mathbf{W_2}$ provides co-phasing between the two polarizations. In the case where $\mathbf{W_1}$ only defines a single beam corresponding to $\mathbf{b}$ being a vector, matrix $\mathbf{W_2}$ would only provide co-phasing between the two polarizations. For transmission ranks $R > 2$ the matrix $\mathbf{W_1}$ defines N_{beams} orthogonal beams where $N_{beams} = \lceil R/2 \rceil$. The N_{beams} beams, together with the two polarization directions in each beam, are then used for transmission of the R layers, with the matrix $\mathbf{W_2}$ only providing co-phasing between the two polarizations. The NR supports transmission of up to eight layers to the same device [14].

In contrast to single-panel CSI, codebooks for Type I multi-panel CSI are designed assuming the joint use of multiple antenna panels at the network side considering that it may be difficult to ensure coherence between transmissions from different panels. More specifically, the design of the multi-panel codebooks assumes an antenna configuration with two or four two-dimensional panels, each with $N_H \times N_V$ cross-polarized antenna elements. The operation principles of Type I multi-panel and single-panel CSI are similar, except that the matrix $\mathbf{W_1}$ defines one beam per polarization and panel, whereas matrix $\mathbf{W_2}$ provides per-subband co-phasing between polarizations as well as panels. The Type I multi-panel CSI supports spatial multiplexing with up to four layers.

Type II CSI provides channel information with significantly higher spatial granularity compared to Type I CSI. Similar to Type I CSI, Type II CSI is based on wideband selection and reporting of beams from a large set of beams. However, while Type I CSI selects and reports a single beam, Type II CSI may select and report up to four orthogonal beams. For each selected beam and each of the two polarizations, the reported PMI then provides an amplitude value (partly wideband and partly subband) and a phase value (subband). This allows constructing a more detailed model of the channel, capturing the main rays and their

respective amplitudes and phases. At the network side, the PMI received from multiple devices can be used to identify a set of devices with which transmission can be done simultaneously on a set of time/frequency resources, that is, MU-MIMO, and what precoder to use for each transmission. Since Type II CSI targets MU-MIMO scenarios, transmission is limited to a maximum of two layers per device [14].

In order to reduce the overhead due to transmission of reference signals, 3GPP NR has moved away from the notion of continuously transmitted wideband cell-specific reference signals and instead has defined more flexible and configurable UE-specific reference signals transmitted on-demand. This transition has resulted in the definition of new UE measurement procedures based on the CSI-RSs. For this purpose, new metrics have been defined that will be explained in detail in the following:

- *Synchronization Signal-Reference Signal Received Power (SS-RSRP)* is defined as the linear average over the power contributions [in (Watts)] of the resource elements that carry secondary synchronization signals. The measurement time resource(s) for SS-RSRP are confined within SS/PBCH block measurement time configuration (SMTC)[9] window size. If SS-RSRP is used for L1-RSRP as configured by reporting configurations, the measurement time resource(s) restriction by SMTC window size is not applicable. For SS-RSRP calculation, DM-RSs for PBCH and, if indicated by higher layers, CSI-RSs in addition to the secondary synchronization signals may be used. The SS-RSRP using DM-RS for PBCH or CSI-RS are measured by linear averaging over the power contributions of the resource elements that carry corresponding reference signals by considering the power scaling for the reference signals. If SS-RSRP is not used for L1-RSRP, the additional use of CSI-RSs for SS-RSRP determination is not applicable. The SS-RSRP is measured only over the reference signals corresponding to SS/PBCH blocks with the same SS/PBCH block index and the same physical-layer cell identity. If SS-RSRP is not used for L1-RSRP and higher layers indicate certain SS/PBCH blocks for performing SS-RSRP measurements, then SS-RSRP is measured only from the indicated set of SS/PBCH block(s). For frequency range 1, the reference point for the SS-RSRP is the antenna connector of the UE. For frequency range 2; however, the SS-RSRP is measured based on the combined signal from antenna elements corresponding to a given receiver branch. For frequency ranges 1 and 2 if receiver diversity is used by the UE, the reported SS-RSRP value must not be lower than the corresponding SS-RSRP of any of the individual receiver branches [10].

[9] SS block-based RRM measurement timing configuration or SMTC is the measurement window periodicity/duration/offset information for UE RRM measurement per carrier frequency. For intra-frequency connected mode measurement, up to two measurement window periodicities can be configured. For the idle mode measurements, a single SMTC is configured per carrier frequency. For inter-frequency connected mode measurements, a single SMTC is configured per carrier frequency.

Table 4.13: NR carrier received signal strength indicator measurement symbols [10].

OFDM Signal Indication *SS-RSSI-MeasurementSymbolConfig*	OFDM Symbol Indices
0	{0,1}
1	{0,1,2,. . .,10,11}
2	{0,1,2,. . .,5}
3	{0,1,2,. . .,7}

- *Secondary Synchronization Signal-Reference Signal Received Quality (SS-RSRQ)* is defined as the ratio of $N_{RB} \times SS\_RSRP/NR\_carrier\_RSSI$, where N_{RB} is the number of resource blocks in the received signal strength indicator (RSSI) measurement bandwidth of the NR carrier. The latter measurements are conducted over the same set of resource blocks. The NR carrier RSSI, *NR_carrier_RSSI*, comprises the linear average of the total received power [in (Watts)] observed over certain OFDM symbols within measurement time resource(s), in the measurement bandwidth, over N_{RB} number of resource blocks from all sources, including co-channel serving and non-serving cells, adjacent channel interference, and thermal noise. The measurement time resource(s) for *NR_carrier_RSSI* are confined within SMTC window duration. If indicated by higher layers, for a half-frame with SS/PBCH blocks, *NR_carrier_RSSI* is measured over OFDM symbols of the indicated slots shown in Table 4.13; otherwise, if measurement gap is not used, *NR_carrier_RSSI* is measured over OFDM symbols within SMTC window duration and, if measurement gap is used, *NR_carrier_RSSI* is measured over OFDM symbols corresponding to overlapped time span between SMTC window duration and minimum measurement time within the measurement gap [10]. If higher layers indicate certain SS/PBCH blocks for performing SS-RSRQ measurements, then SS-RSRP is measured only over the indicated set of SS/PBCH block(s). For frequency range 1, the reference point for the SS-RSRQ is the antenna connector of the UE. For frequency range 2, *NR_carrier_RSSI* is measured based on the combined signal from antenna elements corresponding to a given receiver branch, where the combining function for *NR_carrier_RSSI* is the same as the one used for SS-RSRP measurements. For frequency range 1 and 2, if receiver diversity is used by the UE, the reported SS-RSRQ value must not be lower than the corresponding SS-RSRQ of any of the individual receiver branches [10].
- *CSI-Reference Signal Received Power (CSI-RSRP)* is defined as the linear average over the power contributions (in Watts) of the resource elements that carry CSI-RSs configured for RSRP measurements within the measurement frequency region in the predefined CSI-RS occasions. For CSI-RSRP calculation, the CSI-RSs transmitted on antenna port 3000 or antenna ports 3000, 3001 are used. For frequency range 1, the reference point for the CSI-RSRP is the antenna connector of the UE. For frequency range 2;

however, the CSI-RSRP is measured based on the combined signal from antenna elements corresponding to a given receiver branch. For frequency ranges 1 and 2, if receive diversity is used by the UE, the reported CSI-RSRP value must not be lower than the corresponding CSI-RSRP of any of the individual receiver branches [10].

- *CSI-Reference Signal Received Quality (CSI-RSRQ)* is defined as the ratio of $N_{RB} \times CSI\_RSRP/CSI\_RSSI$, where N_{RB} denotes the number of resource blocks used in the CSI-RSSI measurement bandwidth. The latter measurements are conducted over the same set of resource blocks.
- *CSI-Received Signal Strength Indicator (CSI-RSSI)* is the linear average of the total received power (in Watts) observed over the OFDM symbols within measurement time resource(s), in the measurement bandwidth, over N_{RB} number of resource blocks from all sources, including co-channel serving and non-serving cells, adjacent channel interference, and thermal noise. The measurement time resource(s) for CSI-RSSI corresponds to OFDM symbols containing configured CSI-RS occasions. For CSI-RSRQ calculation, CSI-RSs transmitted on antenna port 3000 are used. For frequency range 1, the reference point for the CSI-RSRQ is the antenna connector of the UE, whereas for frequency range 2, CSI-RSSI is measured based on the combined signal from antenna elements corresponding to a given receiver branch, where the combining for CSI-RSSI is the same as the one used for CSI-RSRP measurements. For frequency ranges 1 and 2, if receive diversity is used by the UE, the reported CSI-RSRQ value must not be lower than the corresponding CSI-RSRQ of any of the individual receiver branches [10].
- *Synchronization Signal-Signal-to-Interference Plus Noise Ratio (SS-SINR)* is defined as the linear average over the power contribution (in Watts) of the resource elements carrying secondary synchronization signals divided by the linear average of the noise and interference power contribution (in Watts) over the resource elements carrying secondary synchronization signals within the same frequency region. The measurement time resource(s) for SS-SINR are confined within SMTC window duration. For SS-SINR calculation, the DM-RSs associated with PBCH in addition to the secondary synchronization signals may be used. If the RRC signaling identifies certain SS/PBCH blocks for conducting SS-SINR measurements, then SS-SINR is measured only over the set of SS/PBCH block(s) identified via signaling. For frequency range 1, the reference point for the SS-SINR is the antenna connector of the UE, whereas for frequency range 2, the SS-SINR is measured based on the combined signal from antenna elements corresponding to a given receiver branch. For frequency ranges 1 and 2, if receiver diversity is used by the UE, the reported SS-SINR value must not be lower than the corresponding SS-SINR of any of the individual receiver branches [10].
- *CSI-SINR* is defined as the linear average over the power contribution (in Watts) of the resource elements carrying CSI-RSs divided by the linear average of the noise plus interference power contributions (in Watts) over the resource elements carrying CSI-RSs within the same frequency region. For CSI-SINR calculation, the CSI-RSs

transmitted on antenna port 3000 are used. For frequency range 1, the reference point for the CSI-SINR is the antenna connector of the UE, whereas for frequency range 2, the CSI-SINR is measured based on the combined signal from antenna elements corresponding to a given receiver branch. For frequency ranges 1 and 2, if receive diversity is used by the UE, the reported CSI-SINR value must not be lower than the corresponding CSI-SINR of any of the individual receiver branches [10].

- *SRS-RSRP* is defined as linear average of the power contributions (in Watts) of the resource elements carrying the uplink SRS. The SRS-RSRP is measured over the configured resource elements within the measurement frequency region and over the time resources in the predefined measurement occasions. For frequency range 1, the reference point for the SRS-RSRP is the antenna connector of the UE. If receive diversity is used by the UE, the reported SRS-RSRP value must not be lower than the corresponding SRS-RSRP of any of the individual receiver branches [10].

The NR supports operating frequencies in a wide range from 450 MHz to 52.6 GHz (and higher in the future releases). The main challenge in NR is to overcome the higher propagation loss and sensitivity to blockage in the above 6 GHz frequency bands. To overcome this issue, efficient usage of highly directional beamformed transmission and reception using a larger number of antenna elements is crucial at the gNB and the UE. To achieve large beamforming gain with reasonable implementation complexity, hybrid beamforming was found to be a suitable solution. The analog beams on each panel/subarray are adapted through phase shifters and/or amplitude scaling. Digital beamforming is adapted by applying different digital precoders across panels/subarrays. At the gNB, downlink transmission with analog beamforming, that is, transmitter beam pointing toward a certain direction can only cover a limited area due to its relatively narrow beam-width. Therefore, the gNB needs to utilize multiple transmit beams to cover the entire cell. Similarly, in the uplink the gNB needs to utilize multiple receive beams to receive the uplink transmissions from the entire cell.

4.1.6.2 Beam Management

The new radio provides a set of mechanisms by which the UEs and the gNB can establish highly directional transmission links, typically using large-scale phased arrays, to benefit from the resulting beamforming gain and to sustain an acceptable communication quality. Directional links, nevertheless, require fine alignment of the transmitter and receiver beams that can be only achieved through a set of procedures known as beam management. The beam management procedures are essential to perform a variety of radio access network functions including the initial access for idle users, which allows a mobile UE to establish a physical connection with a gNB and beam tracking for connected users, which enables beam adaptation schemes, or handover, path selection and radio link failure (RLF) recovery procedures. In LTE, these control procedures are performed using omnidirectional

transmission, and beamforming or other directional transmissions can only be performed after a physical link (user-plane) is established. In certain conditions such as operation in the mmWave bands, it may be necessary to exploit high antenna gains even during initial access and in general for improving the control channel coverage. However, directionality can significantly delay the access procedures and make the performance more sensitive to the beam alignment.

The Rel-15 NR is designed to support analog beamforming in addition to digital precoding/beamforming. In high frequencies, analog beamforming may be beneficial from an implementation viewpoint despite the fact that analog beamforming may constrain the transmit/receive beam to be formed in one direction at a given time and further requires beam sweeping, where the same signal is repeated in multiple OFDM symbols but on different transmit beams. Beam sweeping would ensure that the signal can be transmitted with high directionality in order to cover the entire cell area. The NR has specified control/signaling schemes to support beam management procedures including an indication to the device to assist the selection of an appropriate receive beam. For large number of antennas, beams are narrow and beam tracking may fail; therefore, beam-recovery procedures have been defined where a device can trigger the beam-recovery procedure. A cell may have multiple transmission points each with beams and the beam-management procedures to allow UE-transparent mobility for seamless handover between the beams of different transmission points. In addition, uplink-centric and reciprocity-based beam management is also possible by utilizing uplink signals [14]. In some cases, a suitable transmit/receive beam pair for the downlink transmission will also be a suitable beam pair for the uplink transmission direction and vice versa. The NR refers to this as downlink/uplink beam correspondence, where it is sufficient to determine a suitable beam pair in one direction and use the same beam pair in the opposite direction. Since beam management is not intended to track fast-varying and frequency-selective channels, beam correspondence does not require that downlink and uplink transmissions to take place on the same carrier frequency; thus the concept of beam correspondence is also applicable to FDD systems in paired spectrum.

The beam management is defined as the process of acquiring and maintaining a set of beams, which are originated at the gNB and/or the UE and can be used for downlink and uplink transmission and reception. The beam management process comprises the following functions as shown in Fig. 4.43 [11]:

- *Beam sweeping:* Covering a spatial area with a set of beams transmitted and received according to prespecified intervals and directions. The measurement process is carried out with an exhaustive search, that is, both UE and the base station have a predefined codebook of directions (each identified by a beamforming vector) that cover the entire angular space and are used sequentially to transmit/receive synchronization and reference signals.

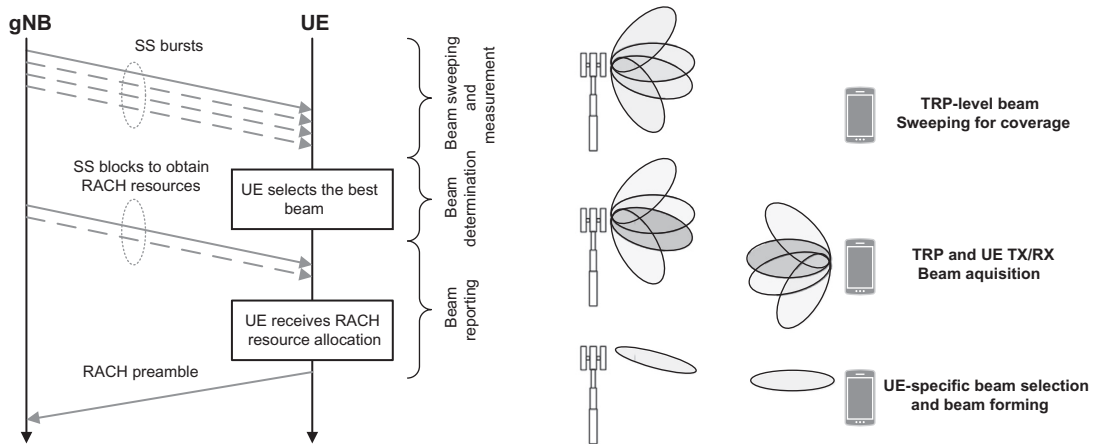


Figure 4.43
Signals and messages exchanged during downlink beam management procedure [69].

- *Beam measurement:* Evaluation of the quality of the received signal at the gNB or at the UE. Different metrics may be used for this purpose such as SNR, which is the average of the received power on synchronization signals divided by the noise power. The measurements for the initial access are based on the SS blocks. The tracking is done using the measurements conducted on the SS bursts and the CSI-RSs, which include a set of directions that may cover the entire set of available directions based on the UE requirements.
- *Beam determination:* Selection of the suitable beam or beams either at the gNB or at the UE, according to the measurements obtained via the beam measurement procedure. This process further allows the TRP(s) or the UEs to select their own transmit/receive beam(s). In beam determination, the gNB and the UE find a beam direction to ensure good radio link quality for the unicast control and data channel transmission. Once a link is established, the UE measures the link quality of multiple transmit and receive beam pairs and reports the measurement results to the gNB. Furthermore, the UE mobility, orientation, and channel blockage can change the radio link quality of the transmitted and received beam pairs. When the quality of the current beam pair degrades, the gNB and the UE can switch to another beam pair with better radio link quality. The gNB and the UE can monitor the quality of the current beam pair along with some other beam pairs and perform beam switching when necessary. When the gNB assigns a transmit beam to the UE via downlink control signaling, the beam indication procedure is used.
- *Beam reporting:* The procedure used by the UE to send beam quality and beam decision information to the gNB where the UE reports its observation of beamformed signal(s) based on beam measurement to the gNB. For the initial access, after beam

determination, the UE must wait for the gNB to schedule the random-access channel (RACH) opportunity corresponding to the best direction that the UE has determined, for performing the random access and implicitly informing the selected serving infrastructure of the optimal direction (or set of directions) through which it must steer its beam, in order to be properly aligned with the UE. As we mentioned earlier, with each SS block, the gNB will specify one or more RACH opportunities with a certain time and frequency offset and direction, so that the UE knows when to transmit the RACH preamble. This may require an additional complete directional scan of the gNB, thus further increasing the time it takes to access the network. For beam tracking in the connected mode, the UE can provide feedback using the control channel that it has already established, unless there is a link failure and no directions can be recovered using CSI-RS. In this case, the UE must repeat the initial access procedure or try to recover the link using the SS block bursts while the user experiences a service unavailability.

- *Beam switching and recovery:* Beam recovery involves a procedure when the link between the gNB and the UE can no longer be maintained and needs to be reestablished.

These procedures are periodically repeated to update the optimal transmitter and receiver beam pair over time. There are two network deployment scenarios for NR, that is, non-standalone and standalone, which can affect the way these procedures are performed. In non-standalone scenario, an NR gNB uses an LTE cell as an anchor for the control-plane management and mobile terminals exploit multi-connectivity to maintain multiple connections to different cells so that any link failure can be overcome by switching data paths. However, such an option may not be available in standalone deployments. The beam management procedure is further illustrated in Fig. 4.44.

The beam management operation in general is based on the control messages which are periodically exchanged between transmit and receive nodes. The reference signals used for beam management depend on the state of the UE. In idle mode, the PSS, SSS, and PBCH DM-RS are used, whereas in the connected mode, the CSI-RS and SRS are used in the downlink and uplink, respectively. A radio connection between a gNB with $N_{gNB-beam}$ analog beams and a UE with $N_{UE-beam}$ analog beams has a total of $N_{gNB-beam} \times N_{UE-beam}$ TX-RX beam pairs. Given that the number of TX/RX beams is typically large in mmWave bands to achieve sufficient coverage, efficient beam measurement and reporting procedures are important to ensure minimal overhead and UE complexity. A gNB capable of transmitting $N_{gNB-beam}$ analog beams can configure up to N_{RS} reference signals for beam measurement. Each RS is beamformed with its associated analog beam pointing in a particular direction. The analog beam associated with each reference signal may also be kept fixed over certain time intervals to allow the UE to test different RX beams for a given TX beam. In this manner, up to $N_{gNB-beam} \times N_{UE-beam}$ beam pairs can be measured.

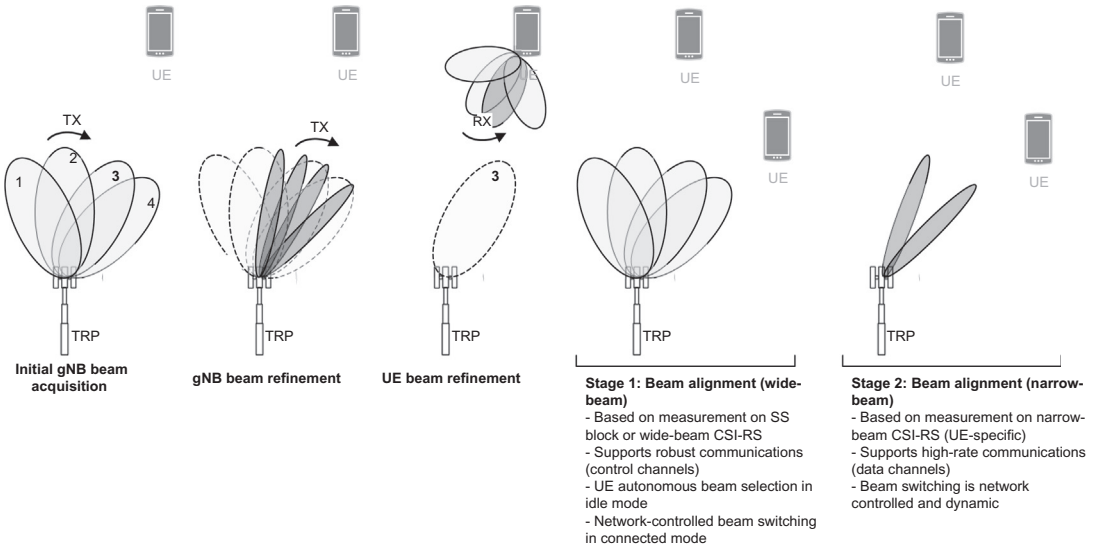


Figure 4.44
Illustration of beam management procedure [26].

The beamformed CSI-RS used for beam management can be transmitted either aperiodically or periodically. When the system is underloaded, beam sweeping across a small number of TX beams over a narrow angular area on an aperiodic basis for a given UE is sufficient. When the system loading increases, it would be more efficient to perform a periodic sweep over a wider angular area covering a larger number of UEs. Despite the fact that $N_{gNB-beam} \times N_{UE-beam}$ beam pairs correspond to $N_{gNB-beam} \times N_{UE-beam}$ beam qualities along with their pair indicators, not all $N_{gNB-beam} \times N_{UE-beam}$ beam qualities need to be reported, because for each TX beam, only the measurement quality of one beam pair, that is, the optimal RX beam for the given TX beam, needs to be reported. The optimal RX beam is known at the UE based on the beam measurement, which is not reported. In the subsequent data or control transmission to the UE, the gNB indicates the index of the selected TX beam to the UE. The UE can use the latest optimal RX beam of the indicated TX beam, which is stored in the UE local memory upon beam measurement, for the purpose of reception. Moreover, within the $N_{gNB-beam}$ candidate TX beams, the qualities and/or indices of $1 \leq n_{beam} \leq N_{gNB-beam}$ TX beams can be reported. If $n_{beam} = 1$, the UE may report the optimal TX beam index without the associated beam quality as a recommendation to the gNB for downlink beamforming. If $n_{beam} > 1$, the UE can report the indices of n_{beam} selected TX beams, for example, the best n_{beam} beams, along with their measured relative/absolute qualities to the gNB. The gNB compares the qualities of the reported TX beams and selects one TX beam for downlink transmission. The NR also supports lower complexity beam management procedures that do not require explicit indication of the selected TX beam to the UE. If the UE only reports the best TX beam index and

only a single gNB TX-RX beam pair is used for transmission of both data and control, there is no need for explicit beam indication. In that case, the UE identifies the optimal RX beam for a given measurement, and the gNB uses the TX beam recommended by the UE for the subsequent data and control transmissions. To receive the data and control transmissions, the UE uses the RX beam that it identified in the previous measurement. The beam quality metrics that can be used for beam measurement are for example RSRP, RSRQ, and SINR calculated based on CSI-RS. Different metrics result in different UE complexity and may be suitable for different scenarios, for example, the RSRP measurement is relatively simple and more power efficient and allows fast measurement of a large number of beams, which is useful for initial beam acquisition. The CSI measurement is more complex but offers more accurate beam information, which can be used for beam refinement within a small group of candidate beams [57].

Following the initial access and connection set up, the device can assume that network transmissions to the device will use the same transmit beam that was used for SS block acquisition. Therefore, the device can assume that the receive beam which was used to acquire the SS block will also be a suitable beam for the reception of subsequent downlink transmissions. Similarly, subsequent uplink transmissions would use the same beam that was used for the random-access preamble transmission, implying that the network can assume that the uplink receive beam established at the initial access will remain valid.

In LTE, a UE continuously performs radio link monitoring of the channel quality of its serving cell to ensure sufficient coverage of control channels. If the link quality is considered poor, the UE declares RLF and triggers a higher layer connection reestablishment procedure, resulting in cell reselection. For a large number of antennas, the UE-specific beams are narrow, and beam tracking can fail, for example, when a moving object blocks the LoS path to the UE. This event is regarded as beam failure in NR. In this case declaring RLF and performing cell reselection is unnecessary since another beam from the same cell can be used to cover the UE. This physical layer procedure is an example of beam recovery, in which the UE continuously monitors a UE-specific periodic reference signal associated with the TX beam with which a control channel is transmitted. If the measured beam quality deteriorates, the UE declares beam failure and proceeds to identify an alternative candidate TX beam, selected from a set of UE-common TX beams used for the periodic beam sweeping of initial access signals. These beams are typically wider compared to UE-specific data beams. As a part of this TX beam determination, the UE also determines an appropriate RX beam to receive the indicated TX beam. When a new beam is detected, the UE transmits a beam recovery request message using a preconfigured uplink resource (which includes an identifier of the new beam) to the serving cell. The network then transmits a recovery response to the UE. If the response is successfully received by the UE, the beam recovery procedure is successful, and a new beam pair is established; otherwise, the UE may perform additional beam recovery attempts, which upon failure would force the UE to initiate the RLF procedure, which includes cell reselection [57].

The beamforming weights for data transmission are typically obtained from codebook-based CSI reporting. Once the codebook is specified, the beamforming capability in azimuth and elevation dimension is restricted by the codebook size, parameters, and structure. Alternatively, non-codebook-based beamforming can be extended to enable a gNB to reuse beam management procedures for acquiring the beamforming weights. In particular, a UE is configured with $K > 1$ CSI-RS resources, each associated with a TX beam. The beamforming can be performed in either digital or analog domain. However, in lower carrier frequencies, digital beamforming may be predominantly utilized. The UE measures the configured K CSI-RS resources and selects $N_{CSI} \leq K$ CSI-RS resources based on their respective qualities and reports the N_{CSI} CRIs. The actual CSI (e.g., CQI/PMI/RI) measured from the selected CSI-RS resources are also reported together with CRIs. As such, beam reporting and CSI reporting can be combined to acquire CSI. When a certain degree of channel reciprocity is achievable, beamforming weights can be derived at the gNB by measuring uplink signal(s) transmitted from the UE. The CSI reporting from the UE can also be utilized by the gNB to determine the beamforming weights [57].

To enable analog beamforming at the UE receiver, different reference signals within the resource set should be transmitted in different symbols, allowing the receiver-side beam to sweep over the set of reference signals. At the same time, the device can assume that different reference signals in the resource set are transmitted using the same spatial filter or alternatively the same transmit beam. In general, a configured resource set includes a repetition flag that indicates whether a device can assume that all reference signals within the resource set are transmitted using the same spatial filter. For a resource set to be used for downlink receiver-side beam adjustment, the repetition flag should be set [9].

4.1.7 Channel Coding and Modulation Schemes

Channel coding is one of the areas where NR is taking a completely different approach from LTE. In NR, LDPC coding has replaced the turbo coding that was previously used for LTE PDSCH/PUSCH coding and polar codes have replaced the tail biting convolutional codes used previously for LTE PDCCH/PUCCH/PBCH coding, except for very small block lengths where repetition/block coding may be used. Turbo codes generally have a low encoding complexity and high decoding complexity, whereas LDPC codes have more complex encoding and less complex decoding algorithms. Considering eMBB use cases with large code block sizes and the code rates up to 8/9, turbo codes may not meet the implementation complexity required for the decoder. The LDPC codes, on the other hand, have relatively simple and practical decoding algorithms. The decoding is performed by iterative belief propagation (BP). The accuracy of decoding will be improved in each iteration, and the number of iterations is decided based on the requirement of the application, providing a trade-off between the bit error performance,

latency, and complexity. In terms of latency, LDPC codes are parallel in nature, while turbo codes are serial in nature, allowing the LDPC codes to better support low latency applications than turbo codes. Furthermore, the bit error rate (BER) of the turbo codes have higher error floor compared to that of LDPC, and the LDPC matrix can be extended to lower rates than LTE turbo codes, achieving higher coding gains for low-rate applications targeting high reliability. As for the polar codes, they were introduced in 2009 and they are among the capacity-achieving codes with low encoding and decoding complexity. They provide full flexibility with very good performance with any code length and code rate without error floor, that is, they do not suffer a decrease in the slope of BLER versus SNR.

4.1.7.1 Principles of Polar Coding

In order to describe the polar encoding/decoding concept, let us review a few prerequisites. The symmetric capacity is defined as the highest possible rate that can be achieved when all of the input symbols to the channel are equiprobable. The mutual information of a binary-input discrete memoryless channel (symmetric capacity) with input alphabet $\mathbb{X} = \{0, 1\}$ is defined as follows [29]:

$$I(W) = \sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} \frac{1}{2} W(y|x) \log_2 \frac{2W(y|x)}{W(y|0) + W(y|1)}$$

The symmetric capacity is equal to the Shannon capacity when the channel W is a symmetric channel. The Bhattacharyya parameter $Z(W)$ is the upper bound on the probability of a maximum likelihood decision error when transmitting 0 or 1 over the channel W. Thus the Bhattacharyya parameter $Z(W)$ is a channel reliability measure. The Bhattacharyya parameter can be calculated as follows:

$$Z(W) = \sum_{y \in \mathbb{Y}} \sqrt{W(y|0)W(y|1)}$$

The relationship between $I(W)$ and $Z(W)$ for any binary-input, discrete, memoryless channel W can be described as follows:

$$I(W) \geq \log \frac{2}{1 + Z(W)}, \quad I(W) \leq \sqrt{1 - Z(W)^2}$$

which means $I(W) = 1$ or $I(W) = 0$, if and only if $Z(W) = 0$ or $Z(W) = 1$, respectively.

Polar codes are the first type of forward error correction codes achieving the symmetric capacity for arbitrary binary-input discrete memoryless channel under low-complexity encoding and low-complexity successive cancelation (SC) decoding with order of $\mathcal{O}(N \log N)$ for infinite length codes. Polar codes are founded based on several concepts including channel polarization, code construction, polar encoding, which is a special case of the normal encoding process (i.e., more structural) and its decoding concept [29].

Channel polarization is the first phase of polar coding where N distinct channels are synthesized such that each of these channels is either completely noisy or completely noiseless, that is, strictly valid for infinite code length N. The measure of how much a channel is noisy in the context of the polar codes was first determined by the symmetric capacity or the Bhattacharyya parameter of the channel; however, BER was used later as a common measure.

Code construction phase involves selecting channels in which the information bits are transmitted. In other words, constructing a polar code means using a vector of bit-channel indices that would be used to transmit information. The rest of the bit-channels would have no data and contain the frozen bits. Several code construction algorithms that vary in complexity, precision, and BER performance exist which include evolution of Z-parameters−based code construction algorithm; Monte-Carlo simulation−based construction algorithm; density evolution−based code construction algorithm; Gaussian approximation−based algorithm; and transition probability matrix−based algorithm [29].

Polar codes are a member of the coset[10] linear block code family, where the information bits are multiplied by a submatrix out of the traditional polar generator matrix, and the frozen bits are multiplied by another submatrix. Polar encoding is characterized by its structural manner, in the sense that all parameters are static, independent of the code rate. Different code rates correspond to different number of information bits, while using the same generator matrix. The systematic polar encoding is an extended version of the nonsystematic polar encoding, where the codeword is first non-systematically encoded, bits at frozen bit-channel positions are reset to the values of the frozen bits, and then non-systematically encoded. Systematic encoding provides better performance in terms of BER than non-systematic encoding. However, both have the same BLER performance. There are various polar decoding algorithms including SC, SC list (SCL), SCL with (SCL-CRC), and BP.

The SC decoder is based on the concept of successively decoding bits, where each stage of bit decoding is based on previously decoded bits. It suffers from inter-bit dependence due to its successive nature and thus error propagation. As a standalone decoder for polar codes, it is outperformed by most polar decoders in terms of BER performance. However, it enjoys a

[10] For a subgroup H of a group G and an element x of G, define xH to be the set $\{xh{:}h\in H\}$ and Hx to be the set $\{hx{:}h\in H\}$. A subset of G in the form of xH for some $x\in G$ is said to be a left coset of H and a subset of the form Hx is said to be a right coset of H. For any subgroup H, we can define an equivalence relation $\sim$ by $x\sim y$ if $x = yh$ for some $h\in H$. The equivalence classes of this equivalence relation are exactly the left cosets of H, and an element x of H is in the equivalence class xH. Thus the left cosets of H form a partition of G. It is also true that any two left cosets of H have the same cardinal number, and in particular, every coset of H has the same cardinal number as $eH = H$, where e is the identity element. Thus the cardinal number of any left coset of H has cardinal number the order of H. The same results are true of the right cosets of G and, in fact, one can prove that the set of left cosets of H has the same cardinal number as the set of right cosets of H.

potential for list decoding, because of its sequential hierarchical structure. It was proved that polar codes achieve Shannon capacity of any symmetric binary-input discrete memoryless channel under SC decoding [33,34].

SCL decoder was proposed as an extended version of SC decoder where instead of successively computing hard decisions for each bit, it branches one SC decoder into two parallel SC-decoders at each stage of decision where each branch has its path metric that is continuously updated for each path. It can be shown that a list of size 32 is enough to almost achieve the ML bound.

SCL with CRC decoder is an extension of SCL decoder, where a high-rate CRC code is appended to the polar code, so that the correct codeword is selected among the candidate codewords from the final list of paths. It was observed that whenever an SCL-decoder fails, the correct codeword exists in the list. Therefore, the CRC was proposed as a validity check for each candidate codeword in the list.

In BP decoder, unlike SC-based decoding techniques, there is no inter-bit dependence and thus no error propagation. It does not encounter any intermediate hard decisions. It updates the LLR values iteratively through right-to-left and left-to-right iterations using the same update functions that were used in LDPC domain. For finite length codes, BP decoder outperforms SC decoder in terms of BER performance.

Channel polarization is the concept upon which the polar codes are built. It is the process through which N distinct channels are generated $\mathrm{W}_N^{(i)}: 1 \leq i \leq N$ from N independent copies of a binary-input discrete memoryless channel. The N generated channels are polarized and have mutual information either close to 0 (i.e., noisy channels) or close to 1 (i.e., noiseless channels). The synthesized channels become perfectly noisy/noiseless as N approaches infinity. The process of channel polarization consists of two phases namely channel combining and channel splitting. In the former phase, N distinct channels are created in $n = \log_2 N$ steps, through recursively combining N copies of a binary-input discrete memoryless channel to form a vector channel $\mathrm{W}_N: X^N \rightarrow Y^N$, where N must be an integer power of two. In the second phase the channel W_N is split into N binary-input channels $\mathrm{W}_N^{(i)}: X \rightarrow Y^N \times X^{i-1}, 1 \leq i \leq N$. Uncoded information bits are transmitted over the reliable or noiseless channels with rate 1 and frozen bits are transmitted over the unreliable or noisy channels [33,34].

Polar code construction is the process of selecting the set of K good channels out of N channels over which uncoded information bits will be transmitted. The selection of the information set $\mathbb{A}$ is done in a channel dependent manner. For finite-length polar codes, the synthesized channels are not fully polarized. Bit errors over the quasi-polarized channels are inevitable. Thus the polar code construction phase is critical to obtain the best possible performance. To construct a polar code, the K reliable channels are chosen to minimize the sum of their Bhattacharyya parameter values $\sum_{i \in \mathbb{A}} Z\left(W_N^{(i)}\right)$ in order to minimize the upper

bound on the block error probability of the constructed polar code. For the binary erasure channel, the Bhattacharyya parameter can be calculated using recursive formulas. Thus the polar code construction problem can be solved without a need for approximation. The Bhattacharyya parameter can be calculated using the following recursive formulas with a complexity of $\mathcal{O}(N)$ [29]:

$$Z(W_N^{(2j-1)}) = 2Z(W_{N/2}^{(j)}) - Z(W_{N/2}^{(j)})^2$$
$$Z(W_N^{(2j)}) = Z(W_{N/2}^{(j)})^2$$
$$Z(W_1^{(1)}) = \epsilon$$

For the AWGN channel no efficient algorithm for calculating the Bhattacharyya parameter per synthesized channel is known. Approximating the exact polar code construction is possible to reduce complexity by calculating an estimate of the Bhattacharyya parameter per synthesized channel. Several suboptimal construction methods were proposed in the literature with different computational complexities. The main difference between polar codes and Reed−Muller codes is the choice of the information set $\mathbb{A}$. In the case of Reed−Muller codes, the indices of the highest weight rows of the generator matrix $\mathbf{G}$ are selected to carry the information. A polar code of length $N = 2^n$ is generated using generator matrix $\mathbf{G}$ of size $N \times N$. A block of length N, consisting of $N - K$ frozen bits and K information bits, is multiplied by $\mathbf{G}$ to produce the polar codeword $\mathbf{x} = \mathbf{uG}$. The generator matrix $\mathbf{G}$ is based on a kernel that is used to construct the code, where

$$\mathbf{G} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{\otimes \log_2 N}$$

in which $\otimes$ denotes the Kronecker product.[11] A polar encoding lattice, equivalent to $\mathbf{G}$, can also be used as a polar encoder, as shown in Fig. 4.45. Note that successive graph representations have recursive relationships. More specifically, the graph representation for a polar encoding kernel operation having a kernel block size of $N = 2$ comprises a single stage, containing a single XOR. The first of the $N = 2$ kernel encoded bits is obtained as the XOR of the $N = 2$ kernel information bits, while the second kernel encoded bit is equal to the second kernel information bit. For greater kernel block sizes N, the graph representation may be considered to be a vertical concatenation of two graph representations for a kernel block size of $N = 2$, followed by an additional stage of XORs, as shown in Fig. 4.45. In analogy with the $N = 2$ kernel described above, the first $N = 2$ of the N kernel encoded

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

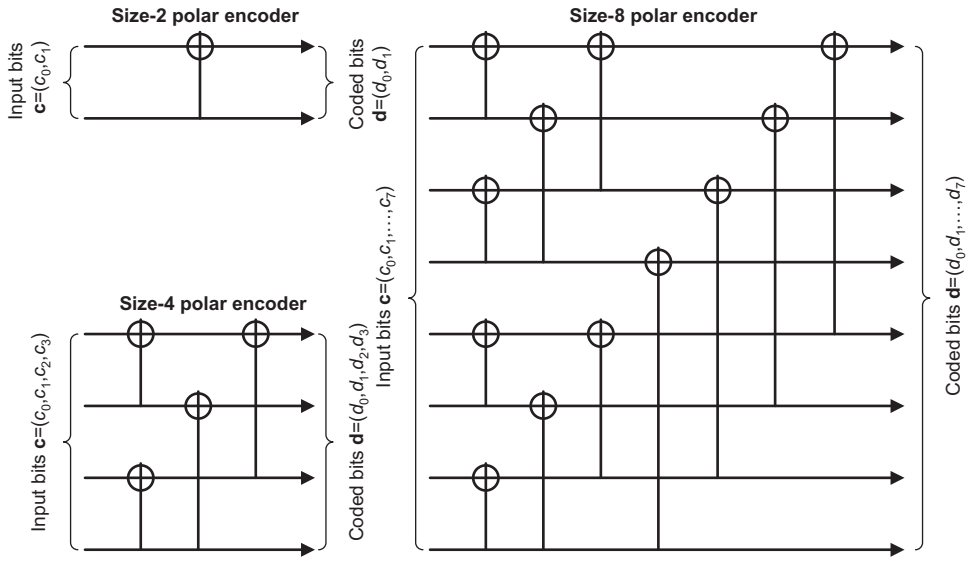


Figure 4.45
Polar encoder of different sizes [32].

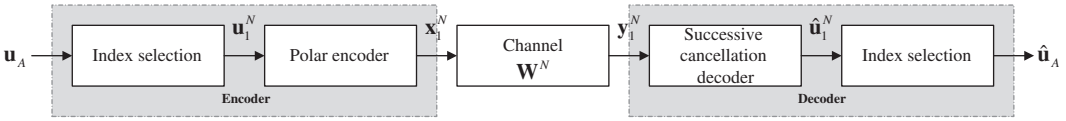


Figure 4.46
High-level polar code encoding and SC list decoding [40].

bits are obtained as XORs of corresponding bits from the outputs of the two $N = 2$ kernels, while the second $N = 2$ of the kernel encoded bits are equal to the output of the second $N = 2$ kernel [35].

Polar codes were introduced as non-systematic codes. Any linear code can be transformed from non-systematic to a systematic code. The systematic polar encoding can be performed using the standard non-systematic polar encoding apparatus in a three-phase operation as follows (see Fig. 4.46):

1. The vector $\mathbf{u} = (u_{\mathbb{A}}, u_{\mathbb{A}^c})$ is encoded in the standard non-systematic fashion producing the vector $\overline{\mathbf{u}}$.
2. The frozen bit positions $\mathbb{A}^c$ in the vector $\overline{\mathbf{u}}$ are set to zero $\overline{\mathbf{u}}_{\mathbb{A}^c} = 0$ (the frozen bits are always set to zero here).
3. The modified vector $\overline{\mathbf{u}}$ is encoded in the standard non-systematic fashion producing the codeword $\mathbf{x}$ which is a systematic polar codeword, in the sense that the information bits $\mathbf{u}_{\mathbb{A}}$ appear in the final codeword $\mathbf{x}$ in the information bits position $\mathbf{x}_{\mathbb{A}}$.

The main advantage of the systematic polar codes is that its BER performance is better than non-systematic polar codes. However, both systematic and non-systematic polar codes have the same BLER performance. It is observed that the systematic polar coding is more robust to error propagation using SC decoder when compared to that in the non-systematic polar coding. There are two main methods for polar decoding namely SC decoder and its variants and BP decoder. The SC decoder and its variants have serial decoding characteristics which cannot be parallelized; thus these decoders suffer from long decoding latency and low throughput, making them not suitable for high-speed applications. The BP decoder can use parallel processing and thus can enhance its throughput, making it suitable for high-speed applications.

In the receiver, the role of the demodulator is to recover information pertaining to the encoded block. However, the demodulator is typically unable to obtain absolute confidence about the value of the bits in the encoded block due to the random nature of the noise in the communication channel. The demodulator may express its confidence about the values of the bits in the encoded block by generating a soft encoded block, which comprises N number of encoded soft bits. Each soft bit may be represented in the form of an LLR as follows: $LLR = \log\left[p(bit = 0)\right] - \log\left[p(bit = 1)\right]$ where $p(bit = 0)$ and $p(bit = 1)$ are the probabilities that the corresponding bit has a value of "0" and "1", respectively. A positive LLR indicates that the demodulator has greater confidence that the corresponding bit has a value of "0", while a negative LLR indicates greater confidence in the bit value of "1". The magnitude of the LLR corresponds to the confidence level, where an infinite value corresponds to absolute confidence, while a magnitude of zero indicates that the demodulator has no information about the bit value [29].

The polar decoder may operate based on different algorithms including SC decoding and SCL decoding. The SC decoder was the first decoder that was used for polar codes which consists of N decision elements for the N bits of $\hat{\mathbf{u}}_1^N$ (see Fig. 4.46). Each of the decision elements computes the hard-decision output based on the observed channel output y_1^N and the previously decoded bits, for example, the kth decision element would compute $\hat{\mathbf{u}}_k$ using y_1^N and $\hat{\mathbf{u}}_1^{k-1}$. The decision element computes the likelihood ratio as follows [29]:

$$L_i \triangleq \frac{W_N^{(i)}\left(y_1^N, \hat{\mathbf{u}}_1^{i-1}|0\right)}{W_N^{(i)}\left(y_1^N, \hat{\mathbf{u}}_1^{i-1}|1\right)}$$

The hard decision per decision element is generated according to the following rule:

$$\hat{\mathbf{u}}_k = \begin{cases} 0, & L_k \geq 1 \\ 1, & L_k < 1 \end{cases}$$

The decision elements of indices that belong to the set $\mathbb{A}^c$ are set to zero $\hat{\mathbf{u}}_k = \mathbf{0}$, that is, the frozen bit positions and values can be considered as the decoder's prior knowledge.

In the SC decoding process, the value selected for each bit in the recovered information block depends on the sign of the corresponding LLR, which in turn depends on the values selected for all previous recovered information bits. If this approach results in the selection of the incorrect value for a particular bit, then this will often result in propagation of errors in all subsequent bits. The selection of an incorrect value for an information bit may be detected with consideration of the subsequent frozen bits, since the decoder knows that these bits should have zero values. More specifically, if the corresponding LLR has a sign that would imply a value of "1" for a frozen bit, then this suggests that an error may have occurred during the decoding of one of the preceding information bits. However, in the SC decoding process, there is no opportunity to consider alternative values for the preceding information bits. Once a value has been selected for an information bit, the SC decoding process is final. This motivates SCL decoding, which enables a list of alternative values for the information bits to be considered. As the decoding process continues, it considers both options for the value of each successive information bit. More specifically, an SCL decoder maintains a list of candidate kernel information blocks, where the list and the kernel information blocks are built up as the SCL decoding process proceeds. At the start of the process, the list comprises only a single-kernel information block having a length of zero bits. Whenever the decoding process reaches a frozen bit, a bit value of "0" is appended to the end of each candidate kernel information block in the list. However, whenever the decoding process reaches an information bit, two replicas of the list of candidate kernel information blocks is created. The bit value of "0" is appended to each block in the first replica and the bit value of "1" is appended to each block in the second replica. Following this, the two lists are merged to form a new list having a length which is double the length of the original list. This continues until the length of the list reaches a limit L, that is, the list size, which is typically a power of two. From this point onwards, each time the length of the list is doubled when considering an information bit, the worst L among the $2L$ candidate kernel information blocks are identified and pruned from the list. Thus the length of the list is maintained at L until the SCL decoding process is completed. In this process, the worst candidate kernel information blocks are identified by comparing and sorting appropriate metrics that are computed for each block, based on the LLRs obtained on the left-hand edge of the polar code graph [35,38].

There are several challenges associated with the hardware implementation of polar encoders and, in particular, polar decoders. For example, the complexity of a polar decoder is much greater than that of a polar encoder for three reasons: (1) while polar encoders operate on the basis of bits, polar decoders operate on the basis of the probabilities of bits, which require more memory to store and more complex computations; (2) while polar encoders only have to consider the particular permutation of the information block that they are presented, polar decoders must consider all possible permutations of the information block and must select the one which is the most likely; and (3) while polar encoders only process each

information block once, an SCL polar decoder must process each information block L times in order to achieve sufficiently strong error correction. For these reasons, the latency, hardware resource usage, and power consumption of polar decoders are typically much greater than those of polar encoders. Another challenge in the implementation of the SCL decoding process is imposed by metric sorting. As described earlier, the sorting is required in order to identify and prune the worst L candidate kernel information blocks, among the merged list of $2L$ candidates. One option is to employ a large amount of hardware to simultaneously compare each of the $2L$ candidates with every other candidate, so that the sorting can be completed within a short time. Alternatively, the hardware resource requirement can be reduced by structuring successive comparisons to efficiently reuse intermediate results at the cost of increasing the latency required to rank the $2L$ candidates. The CRC bits are employed by the NR polar code in order to facilitate error detection and to improve the error correction capability of the polar decoder. However, there is a trade-off between the error detection capability and the error correction capability. In order to meet the BLER requirements of NR for the control channels, the CRC bits must be handled very carefully, in a manner which is not captured in the NR specifications. In particular, the CRC (and parity check or PC) bits must be decoded as an integral part of the polar decoding process, using an unconventional decoding technique [35].

4.1.7.2 NR Polar Coding

3GPP NR uses a variant of the polar code called distributed CRC (D-CRC) polar code, that is, a combination of CRC-assisted and PC polar codes, which interleaves a CRC-concatenated block and relocates some of the PC bits into the middle positions of this block prior to performing the conventional polar encoding described earlier. This allows a decoder to early terminate the decoding process as soon as any parity check is not successful. The D-CRC scheme is important for early termination of decoding process, because the post-CRC interleaver can distribute information and CRC bits such that partial CRC checks can be performed during list decoding and paths failing partial CRC check can be pruned, leading to early termination of decoding. The post-CRC interleaver design is closely tied to the CRC generator polynomial, thus by appropriately selecting the CRC polynomial, one can achieve better early termination gains and maintain acceptable false alarm rate. The signal flow graph of D-CRC polar encoding and decoding is shown in Fig. 4.47.

In NR, the polar code is used to encode broadcast channel as well as DCI and uplink control information (UCI). Let us denote by N_c the number of control bits that must be transmitted using a code of length E bits. We add L_{CRC} CRC bits to the information bits, resulting in K bits that will be encoded by an (N, K) polar encode with $N = 2^n$. Rate matching is performed to obtain a code of length E and effective rate $R = N_c/E$. To each vector $\mathbf{a} = (a_0, a_1, \ldots, a_{N_c-1})$, containing N_c control information bits to be transmitted, L_{CRC}-bit CRC is attached. The resulting vector $\mathbf{c} = (c_0, c_1, \ldots, c_{K-1})$ comprising $K = N_c + L_{CRC}$ bits

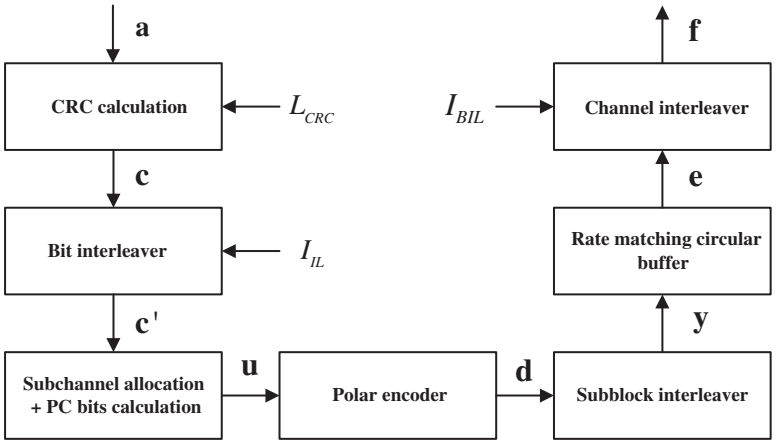


Figure 4.47
3GPP NR polar encoding flow graph [36].

is passed through an interleaver. Based on the desired code rate R and codeword length E, a polar code of length N is utilized along with the relative bit channel reliability sequence and the frozen set. The interleaved vector $\mathbf{c}'$ is assigned to the information set along with the PC bits, while the remaining bits in the N-bit vector $\mathbf{u}$ are frozen. Vector $\mathbf{u}$ is encoded with $\mathbf{d} = \mathbf{uG}$, where the generator matrix $\mathbf{G}$ was defined earlier. After encoding, a subblock interleaver divides $\mathbf{d}$ into 32 equal-length blocks, scrambling them and creating vector $\mathbf{y}$ that is fed into the circular buffer as illustrated in Fig. 4.47. For rate matching, puncturing, shortening, or repetition are applied to change the N-bit vector $\mathbf{y}$ into the E-bit vector $\mathbf{e}$. A channel interleaver is finally applied to compute the vector $\mathbf{f}$ that is now ready to be modulated and transmitted [36].

The NR polar encoder relies on several parameters that depend on the amount and type of information to be transmitted and the physical channel used. The first parameter that needs to be identified is the code length of the polar code, $N = 2^n$. The number n is calculated as $n = \max\{\min\{n_1, n_2, n_{\max}\}, n_{\min}\}$, where $n_{\min}$ and $n_{\max}$ provide a lower and an upper bound on the code length, respectively. In particular, $n_{\min} = 5$ and $n_{\max} = 9$ for the downlink control channel, whereas $n_{\max} = 10$ for the uplink control channel. The parameter $n_2 = \lceil \log_2(K/R_{\min}) \rceil$ gives an upper bound on the code rate based on the minimum code rate admitted by the encoder, that is, 1/8. The value of parameter n_1 is dependent on the rate-matching scheme. It is usually calculated as $n_1 = \lceil \log_2 E \rceil$ so that 2^{n_1} is the smallest power of two larger than E. However, a correction factor is introduced to avoid too severe rate matching: if $\log_2 E < 0.17$, that is, if the smallest power of two larger than E is too far from E, the parameter is set to $n_1 = \lfloor \log_2 E \rfloor$. In this case, an additional constraint on the code dimension is added by imposing $K/E < 9/16$ to ensure that $K < N$. If a code length

Table 4.14: 3GPP NR polar encoding parameters [7].

	PUCCH/PUSCH			PDCCH/PBCH
	$N_c \geq 20$	$12 \leq N_c \leq 19$		
		$E - N_c \leq 175$	$E - N_c > 175$	
$n_{\max}$	10			9
I_{IL}	0			1
I_{BIL}	1			0
L_{CRC}	11	6		24
n_{PC}	0	3		0
n_{PC}^{wm}	0	0	1	0

$N > E$ is selected, the polar code will be punctured or shortened, depending on the code rate before the transmission. In particular, if $K/E \leq 7/16$, the code will be punctured; otherwise it will be shortened. If $N < E$, repetition is used, and some encoded bits will be transmitted twice. In this case, the code construction ensures that $K < N$.

As shown in Table 4.14, a set of parameters are defined to differentiate between different type of control information. The parameters I_{IL} and I_{BIL} refer to the activation of the input bits interleaver and the channel interleaver, respectively. The value of the two types of assistant PC bits are given by n_{PC} and n_{PC}^{wm}. The length of the control information vector N_c and the length of the transmitted codeword E are dependent on the type, content, and number of consecutive transmissions and are reliant on the decisions taken in the higher layers [36].

The K bit output of the CRC encoder is interleaved before being fed to the polar encoder. The interleaver is activated through the I_{IL} flag. In particular, the input bit interleaver is activated for PBCH and PDCCH payloads, while it is set to zero in the case of PUCCH and PUSCH control information. The input bit interleaver interleaves up to $K_{IL}^{\max} = 164$ input bits, where the interleaving pattern is calculated based on the sequence $\Pi_{IL}^{\max}(m)$ given in [7]. The maximum number of input bits $K_{IL}^{\max}$ is set to 164 suggesting that the maximum number of control information bits without CRC is limited to 140. In more detail, $\left(K_{IL}^{\max} - K\right)$ is subtracted from all the entries of $\Pi(k)$, such that $\Pi(k)$ contains the integers smaller than K in permuted order. This scrambling sequence has been proposed to facilitate early termination, both during normal decoding and in DCI blind detection. This is made possible by the fact that after interleaving, every CRC remainder bit is placed after its relevant information bits. The interleaving function is applied to vector $\mathbf{c}$ to obtain the $K-$ bit vector $\mathbf{c}' = (c_{\Pi(0)}, c_{\Pi(1)}, \ldots, c_{\Pi(K-1)})$ [7].

In the subchannel allocation process prior to polar encoding, the vector $\mathbf{c}'$ is expanded into the N-bit vector $\mathbf{u}$ with the addition of assistant bits and frozen bits. As a first step, n_{PC} PC

bits are inserted among the K information and CRC bits. Thus the polar encoder represents a $(N, K + n_{PC})$ code. To create the input vector $\mathbf{u}$ to be encoded, the frozen set of subchannels needs to be identified. The number and position of frozen bits depend on N, E and the selected rate-matching scheme. To begin with, the frozen set $\overline{\mathbf{Q}}_F^N$ and the complementary information set $\overline{\mathbf{Q}}_I^N$ are computed based on the polar reliability sequence $\mathbf{Q}_0^{N_{\max}-1} = \left\{Q_0^{N_{\max}}, Q_1^{N_{\max}}, \ldots, Q_{N_{\max}-1}^{N_{\max}}\right\}$ and the rate matching scheme. The information bits are subsequently assigned to vector $\mathbf{u}$ according to the information set. The assistant PC bits are calculated and stored in $\mathbf{u}$, if necessary [7,36]. The first bits identified in the frozen set correspond to the indices of the $N - E$ bits that are not transmitted, that is, the bits that are punctured from the codeword by the rate matching. These indices correspond to the first $N - E$ or the last $N - E$ codeword bits in the case of puncturing and shortening, respectively. Due to the presence of an interleaver between the encoding and the rate matching, the actual indices to be added to the frozen set correspond to the first or the last bits after the interleaving process. If $K/E \leq 7/16$ and henceforth the polar code must be punctured, additional indices are included in the frozen set to prevent bits in the information set to become ineffective due to puncturing. Furthermore, new indices are added to the frozen set from the reliability sequence starting from the least reliable bits. The polar reliability sequence $\mathbf{Q}_0^{N_{\max}-1} = \left\{Q_0^{N_{\max}}, Q_1^{N_{\max}}, \ldots, Q_{N_{\max}-1}^{N_{\max}}\right\}$ is a list of integers smaller than 1024 sorted in reliability order, from the least reliable to the most reliable; indices larger than N are skipped during the creation of $\overline{\mathbf{Q}}_F^N$.

Unlike the conventional polar encoding process where the Bhattacharyya parameters or in general the reliability factors are calculated prior to encoding process, in 3GPP NR, those reliability factors are tabulated in the standard specification. Prior to encoding, the polar sequence $\mathbf{Q}_0^{N_{\max}-1} = \left\{Q_0^{N_{\max}}, Q_1^{N_{\max}}, \ldots, Q_{N_{\max}-1}^{N_{\max}}\right\}$, in which $0 \leq Q_i^{N_{\max}} \leq N_{\max} - 1$ for $i = 0, 1, \ldots, N_{\max} - 1$ denotes a bit index, is sorted in ascending order of reliability factors $W\left(Q_0^{N_{\max}}\right) < W\left(Q_1^{N_{\max}}\right) < \cdots < W\left(Q_{N_{\max}-1}^{N_{\max}}\right)$, where $W\left(Q_i^{N_{\max}}\right)$ denotes the reliability of bit index $Q_i^{N_{\max}}$. For any code block of length N bits, the same polar sequence $\mathbf{Q}_0^{N-1} = \left\{Q_0^N, Q_1^N, Q_2^N, \ldots, Q_{N-1}^N\right\}$ is utilized. The polar sequence $\mathbf{Q}_0^{N-1}$ is a subset of polar sequence $\mathbf{Q}_0^{N_{\max}-1}$ with all elements $Q_i^{N_{\max}}$ of values less than N, ordered in ascending order of reliability factors $W\left(Q_0^N\right) < W\left(Q_1^N\right) < W\left(Q_2^N\right) < \cdots < W\left(Q_{N-1}^N\right)$. In the preceding expressions, sequence $\overline{\mathbf{Q}}_I^N$ denotes the set of information bit indices in the polar sequence $\mathbf{Q}_0^{N-1}$, and $\overline{\mathbf{Q}}_F^N$ is the set of other bit indices in polar sequence $\mathbf{Q}_0^{N-1}$, where $\overline{\mathbf{Q}}_I^N$ and $\overline{\mathbf{Q}}_F^N$ are derived through subblock interleaving, $\left|\overline{\mathbf{Q}}_I^N\right| = K + n_{PC}$ (i.e., cardinality of the set or the length of the sequence), $\left|\overline{\mathbf{Q}}_F^N\right| = N - \left|\overline{\mathbf{Q}}_I^N\right|$ and n_{PC} is the number of PC bits (see Table 4.15).

Table 4.15: Polar sequence $Q_0^{N_{\max}-1}$ and the associated reliability factor $W\left(Q_i^{N_{\max}}\right)$ [7].

$W\left(Q_i^{N_{\max}}\right)$	$Q_i^{N_{\max}}$	$W\left(Q_i^{N_{\max}}\right)$	$Q_i^{N_{\max}}$	$W\left(Q_i^{N_{\max}}\right)$	$Q_i^{N_{\max}}$	$W\left(Q_i^{N_{\max}}\right)$	$Q_i^{N_{\max}}$	$W\left(Q_i^{N_{\max}}\right)$	$Q_i^{N_{\max}}$	$W\left(Q_i^{N_{\max}}\right)$	$Q_i^{N_{\max}}$	$W\left(Q_i^{N_{\max}}\right)$	$Q_i^{N_{\max}}$	$W\left(Q_i^{N_{\max}}\right)$	$Q_i^{N_{\max}}$
0	0	128	518	256	94	384	214	512	364	640	414	768	819	896	966
1	1	129	54	257	204	385	309	513	654	641	223	769	814	897	755
2	2	130	83	258	298	386	188	514	659	642	663	770	439	898	859
3	4	131	57	259	400	387	449	515	335	643	692	771	929	899	940
4	8	132	521	260	608	388	217	516	480	644	835	772	490	900	830
5	16	133	112	261	352	389	408	517	315	645	619	773	623	901	911
6	32	134	135	262	325	390	609	518	221	646	472	774	671	902	871
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
125	768	253	209	381	539	509	248	637	806	765	914	893	506	1021	1021
126	268	254	284	382	111	510	369	638	427	766	752	894	749	1022	1022
127	274	255	648	383	331	511	190	639	904	767	868	895	945	1023	1023

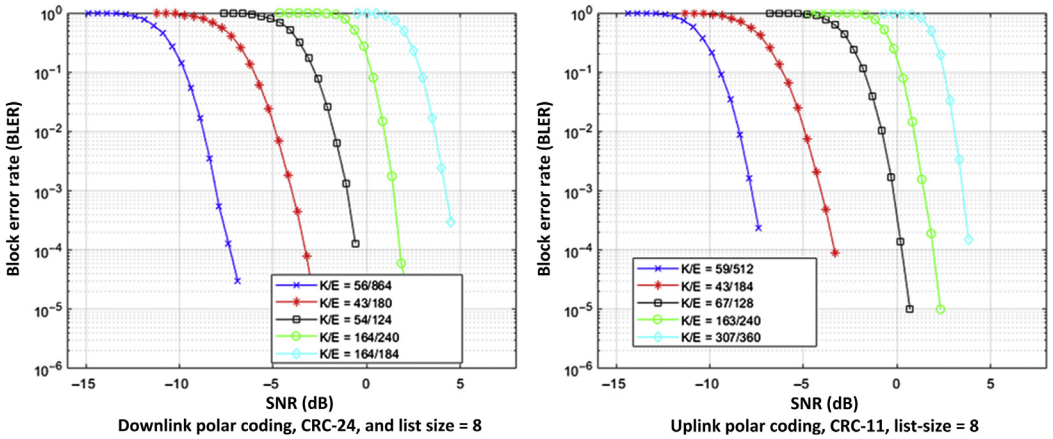


Figure 4.48
Downlink/uplink polar coding with various K/E ratios [67].

Once the input bits are reordered and moved to the reliable positions according to the above reliable bit positions determination procedure, the reordered bits are passed through the polar encoder [7].

As we defined earlier, the polar code generator matrix $\mathbf{G}_N = (\mathbf{G}_2)^{\otimes n}$ is constructed as the nth Kronecker power of matrix $\mathbf{G}_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. For a bit index $j = 0, 1, \ldots, N-1$, let $\mathbf{g}_j$ denote the jth row of $\mathbf{G}_N$ and $w(\mathbf{g}_j)$ the row weight of $\mathbf{g}_j$, where $w(\mathbf{g}_j)$ is the number of ones in $\mathbf{g}_j$. Let us further assume that $\mathbf{Q}_{PC}^N$ is the set of bit indices for PC bits, where the cardinality of the set is $|\mathbf{Q}_{PC}^N| = n_{PC}$. A number of PC bits are placed in the $(n_{PC} - n_{PC}^{wm})$ least reliable bit indices in $\overline{\mathbf{Q}}_I^N$. Other n_{PC}^{wm} PC bits are placed in the bit indices of minimum row weight in $\tilde{\mathbf{Q}}_I^N$, where $\tilde{\mathbf{Q}}_I^N$ denotes the $\left(\left|\overline{\mathbf{Q}}_I^N\right| - n_{PC}\right)$ most reliable bit indices in $\overline{\mathbf{Q}}_I^N$. If there are more than n_{PC}^{wm} bit indices of the same minimum row weight in $\tilde{\mathbf{Q}}_I^N$, the other n_{PC}^{wm} PC bits are placed in the n_{PC}^{wm} bit indices of the highest reliability and the minimum row weight in $\tilde{\mathbf{Q}}_I^N$. The output bit sequence following polar encoding $\mathbf{d} = [d_0, d_1, \ldots, d_{N-1}]$ is obtained as $\mathbf{d} = \mathbf{u}\mathbf{G}_N$ where vector $\mathbf{u} = [u_0, u_1, \ldots, u_{N-1}]$ is derived from interleaved sequence $c'_0, c'_1, \ldots, c'_{K-1}$ following the above bit reordering process. The encoding is performed in $GF(2)$[12].

The performance of the polar codes with different K/E ratios is shown in Fig. 4.48.

[12] $GF(2)$ is the Galois field comprising two elements and the smallest finite field. One may also define $GF(2)$ as the quotient ring of the ring of integers $\mathbb{Z}$ by the ideal $2\mathbb{Z}$ of all even numbers $GF(2) = \mathbb{Z}/2\mathbb{Z}$.

4.1.7.3 Principles of Low Density Parity Check Coding

LDPC codes belong to the class of forward error correction codes which are used for sending a message over noisy transmission channel. These codes can be described by a parity check matrix which contains mostly zeros and a relatively small number of ones. Thus the decoding complexity is small when compared to other code constructions. A very efficient iterative decoding algorithm known as belief propagation is used in the decoder. The LDPC codes can be divided into two groups: regular LDPC codes when the column weight and the row weight of the PC matrix are constant and equal and irregular LDPC codes when the column weight and the row weight are not constant and equal, meaning that the number of ones per row and column is different.

The LDPC codes are represented in different ways. Similar to all linear block codes, a matrix representation by the corresponding generator matrix $\mathbf{G}$ or the PC matrix $\mathbf{H}$ is possible. Thus if the number of input information bits is K and the number of output bits is N, the PC matrix $\mathbf{H}$ is expressed as an $M \times N$ matrix, where $M = N - K$. The resultant code rate K/N defines the size of the PC matrix. The LDPC codes can be further graphically represented with a Tanner graph, which is one of the most common graphical representations for the LDPC codes. It provides the complete representation of the code and helps to describe the decoding algorithm. Tanner graphs are bipartite graphs, that is, there are two disjunct sets of nodes. The two types of nodes are variable nodes (VND) and check nodes (CND). The VNDs represent the code bits, thus, each of the N columns of matrix $\mathbf{H}$ is represented by one VND. The CNDs represent the code constraints; thus each of M rows of matrix $\mathbf{H}$ is represented by one CND. Each VND v_i is connected to a CND c_j, if $h_{ij} = 1$. For the following example PC matrix, the Tanner graph is as shown in Fig. 4.49.

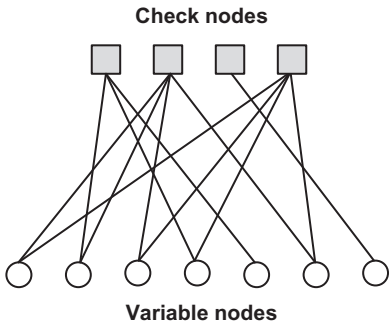


Figure 4.49
Example Tanner graph [38].

$$\mathbf{H} = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Quasi-cyclic (QC) LDPC codes belong to the class of structured codes that are relatively easier to implement without significantly compromising the performance of the code. The QC-LDPC codes can be implemented using simple shift registers with linear complexity based on their generator matrices. Well-designed QC-LDPC codes have been shown to outperform computer-generated random LDPC codes, in terms of bit-error rate and block-error rate performance and the error floor. These codes also have advantages in decoder hardware implementation due to their cyclic symmetry, which results in simple regular interconnection and modular structure. In most of the wireless communication standards including 3GPP NR, a base graph $\mathbf{u}$ is used to define the LDPC code. However, $\mathbf{u}$ needs to be transformed into a PC matrix $\mathbf{H}$ using a lifting factor Z. Lifting means that each (integer) entry of base graph $\mathbf{u}$ is replaced by a permuted $Z \times Z$ identity matrix. We start with an identity matrix $\mathbf{I}$ and circularly shift the entries of this matrix according to the base graph entry u_{ij} to obtain the desired matrix $\mathbf{H}$. As an example, suppose a 2×2 base graph matrix $\mathbf{u}$ and lifting factor $Z = 3$ are given. The transformation from $\mathbf{u}$ to $\mathbf{H}$ can be performed as follows:

$$\mathbf{u} = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \rightarrow \mathbf{H} = \begin{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} & \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} & \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \end{bmatrix} \rightarrow \mathbf{H} = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

The LDPC codes are universally specified by their PC matrices. The PC matrix of a QC-LDPC code is given as an array of sparse circulant matrices of the same size. A circulant matrix is a square matrix in which each row is the cyclic shift of the row above it, and the first row is the cyclic shift of the last row. For a circulant matrix, each column is the downward cyclic shift of the column on its left, and the first column is the cyclic shift of the last column. The row and column weights of a circulant matrix are the same. If the row/column weight is equal to one then the circulant matrix is also a permutation matrix. A circulant matrix is fully characterized by its first row (or first column), which is called the generator of the circulant. For an $m \times m$ circulant matrix $\mathbf{A}$ over $GF(2)$, if its rank $r = m$, then all of its rows are linearly independent. However, if its rank $r < m$ then any consecutive r rows (or columns) of $\mathbf{A}$ may be regarded as linearly independent, and the other $m - r$ rows (or columns) are linearly dependent. This is due to the cyclic structure of $\mathbf{A}$. A QC-LDPC code is given by the null space of an array of sparse circulant matrices of the same size.

An LDPC code can be defined by PC matrix $\mathbf{H}$. For each codeword $\mathbf{v}$, it can be shown that $\mathbf{H}\mathbf{v}^T = \mathbf{0}$. A non-codeword (corrupted codeword) on the other hand, will generate a non-zero vector, which is called syndrome. Mathematically, an LDPC code is the null-space of the PC matrix $\mathbf{H}$. A regular LDPC block code (d_v, d_c) has VND degree d_v and CND degree d_c, which is equal to the column (row)-weight of $\mathbf{H}$. The PC matrix represents a system of linear equations where each row can be represented by a linear combination of the CNDs. Any set of vectors that span the row-space generated by $\mathbf{H}$ can serve as the rows of a PC matrix. The degree of a node is the number of edges (lines) connected to it in a Tanner graph.

The decoding algorithms for LDPC codes were discovered independently, and they come under different names. The most common ones are the BP algorithm, the message passing algorithm (MPA), and the sum-product algorithm. In a BP algorithm, the probabilistic messages are iteratively exchanged between variable and check nodes until either a valid codeword is found or the maximum number of iterations is exceeded. The LDPC codes can be decoded using message passing or BP on the bipartite Tanner graph where, the CNDs and VNDs communicate with each other, successively passing revised estimates of the associated LLR in each decoding iteration. The bit reliability metric is defined as $LLR(b_i) = \log p(b_i = 0) - \log(b_i = 1)$ where b_i denotes the ith bit in the received codeword. If $LLR > 0$, it implies that $b_i = 0$ is more likely, while $LLR < 0$ implies that $b_i = 1$ is more probable. As an example, assume that codeword $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ is transmitted, and $(0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0)$ is received by decoder. Each valid 11-bit codeword $\mathbf{c} = (c_0, c_1, \ldots, c_{10})$ has the sum (modulo 2) of all bits equal to zero. The received vector does not satisfy this code constraint, indicating that there are errors present in the received codeword. Furthermore, assume that the decoder is provided with bit-level reliability metric in the form of probability (confidence in the received values) of being correct as $(0.9, 0.8, 0.86, 0.7, 0.55, 1, 1, 0.8, 0.98, 0.68, 0.99)$. From the soft information, it follows that bit c_4 is the least reliable and should be flipped to bring the received codeword in compliance with code constraint. Using LLRs as messages, the hardware implementation has become much easier when compared to the message passing algorithm. The implementation complexity is further reduced by simplifying the process for updating the CNDs, which is the most complex part of the message passing algorithm. This algorithm is known as the min-sum algorithm. An LDPC decoder can be implemented using serial, parallel, or partially parallel architectures. The performance of the LDPC decoder depends on various factors such as decoder algorithm and architecture, quantization of LLRs and the maximum number of decoding iterations. The maximum number of decoding iterations used for the decoding process determines the data rate and latency of the LDPC decoder. After performing maximum number of decoding iterations, the codeword is then estimated. In order to save decoder power consumption and to decrease the latency, a decoder design that verifies the codeword after each iteration and stops the decoding process when the estimated codeword is correct, is needed. If parity check is satisfied then the codeword is estimated at the beginning of the next iteration and the decoding process is stopped.

4.1.7.4 NR Low Density Parity Check Coding

3GPP NR has taken a different approach to LDPC coding for the downlink and uplink traffic channels. In order to ensure support of a wide range of code rates with sufficient granularity and HARQ-IR, two base graphs with the structures that are explained later in this section have been adopted. Code extension of a PC matrix (lower triangular extension, which includes diagonal-extension as a special case) is used to support HARQ-IR and rate-matching. The 3GPP NR LDPC base graphs consist of five submatrices **A**,**B**,**C**,**D**,**E** as shown in Fig. 4.50. As depicted in the figure, matrix **A** corresponds to the systematic bits; **B** is a square matrix and corresponds to the parity bits. The first or last column of matrix **B** has a weight equal to one. The last row of **B** has a non-zero value and a weight equal to one. If there is a column with weight of one then the remaining columns contain a square matrix such that the first column has weight three. The columns after the weight three column have a dual diagonal structure (i.e., main diagonal and off diagonal elements). If there is no column with weight one, **B** consists of only a square matrix such that the first column has weight three, **C** is a zero matrix, **D** corresponds to single parity check rows, and **E** is an identity matrix for the base graph [55].

The rate matching for the LDPC code uses a circular buffer similar to LTE. The circular buffer is filled with an ordered sequence of systematic bits and parity bits. For HARQ-IR, each RV RV_i is assigned a starting bit location s_i in the circular buffer. For HARQ-IR retransmission of RV_i, the coded bits are read out sequentially from the circular buffer, starting with the bit location s_i. Limited buffer rate matching is further supported. Before code block segmentation, $L_{CRC} = 24$ TB-level CRC bits are attached to the end of the transport block. The value of L_{CRC} was determined to satisfy the probability of misdetection of the TB with $BLER < 10^{-6}$ as well as the inherent error detection of LDPC codes.

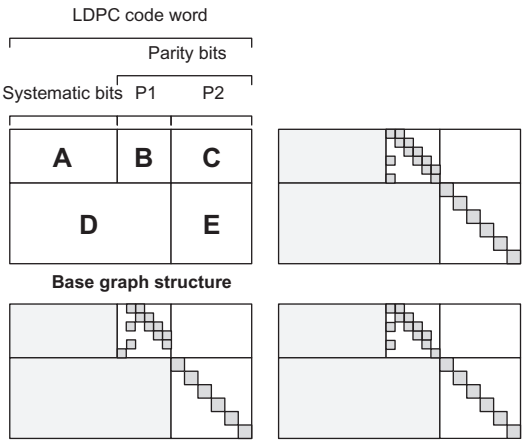


Figure 4.50
Structure of 3GPP NR base graphs with dual diagonal property [55].

The NR LDPC coding chain includes code block segmentation, CRC attachment, LDPC encoding, rate matching, and systematic-bit-priority interleaving. More specifically, code block segmentation allows very large transport blocks (MAC PDUs) to be divided into multiple smaller sized code blocks that can be efficiently processed by the LDPC encoder/decoder. The CRC bits are then attached for error detection purposes. When combined with the inherent error detection of the LDPC codes through the parity check equations, very low probability of undetected errors can be achieved. The rectangular interleaver with number of rows equal to the QAM order improves the performance by making systematic bits more reliable than parity bits for the initial transmission of the code blocks. The NR LDPC codes use a QC structure, where the PC matrix is defined by a smaller base graph. Each entry of the base graph represents either a $Z \times Z$ zero matrix or a shifted $Z \times Z$ identity matrix, where a cyclic shift to the right of each row is applied. Unlike the LDPC codes specified in other wireless technologies, the NR LDPC codes have a rate-compatible structure, which means codewords with different rates can be generated by including a different number of parity bits, or equivalently by using a smaller subset of the full PC matrix. This is especially useful for communication systems employing HARQ-IR for retransmissions. Another advantage of this structure is that for higher rates, the PC matrix and the decoding complexity and latency are smaller. This is in contrast with the LTE turbo codes, which have constant decoding complexity and latency irrespective of the code rate [55].

The NR data channel supports two base graphs to ensure good performance and decoding latency can be achieved for the full range of code rates and information block sizes. The base graph 1 is optimized for large information block sizes and high code rates. It is designed for maximum code rate of 8/9 and may be used for code rates up to 0.95. The base graph 2 is optimized for small information block sizes and lower code rates. The lowest code rate for base graph 2 without using repetition is 1/5. This is significantly lower than that of the LTE turbo codes, which rely on repetition for code rates below 1/3. The NR LDPC codes can also achieve an additional coding gain at low-code rates, which makes them suitable for high reliability scenarios. From decoding complexity perspective, for a given number of input bits, it is beneficial to use base graph 2, since it is more compact and utilizes a larger lifting factor, that is, more parallelism, relative to base graph 1. The decoding latency is typically proportional to the number of non-zero elements in the base graph. Since base graph 2 has much fewer non-zero elements compared to base graph 1 for a given code rate, its decoding latency is significantly lower.

In 3GPP NR, the input bit sequence is represented as $\mathbf{c} = [c_0, c_1, \ldots, c_{K-1}]^T$, where K is the number of information bits to encode. The output LDPC-coded bits are denoted by $d_0, d_1, \ldots, d_{N-1}$ where $N = 66Z_c$ for base graph 1 and $N = 50Z_c$ for base graph 2, where the value of lifting factor Z_c is given in Table 4.7. A code block is encoded by the LDPC encoder based on the following procedure [7]:

1. Find the set with index i_{LS} in Table 4.7 which contains Z_c.
2. Set $d_{k-2Z_c} = c_k, \forall k = 2Z_c, \ldots, K-1$.
3. Generate $N + 2Z_c - K$ parity bits $\mathbf{w} = \left[w_0, w_1, \ldots, w_{N+2Z_c-K+1}\right]^T$ such that $\mathbf{H}\begin{bmatrix}\mathbf{c} & \mathbf{w}\end{bmatrix}^T = \mathbf{0}$.
4. The encoding is performed in $GF(2)$. For base graph 1, matrix $\mathbf{H_{BG}}$ (representing the base graph) has 46 rows and 68 columns. For base graph 2, matrix $\mathbf{H_{BG}}$ has 42 rows and 52 columns. The elements of $\mathbf{H_{BG}}$ matrices are given in [7]. The PC matrix $\mathbf{H}$ is obtained by replacing each element of $\mathbf{H_{BG}}$ with a $Z_c \times Z_c$ matrix such that each element of value 0 in $\mathbf{H_{BG}}$ is replaced by an all zero matrix $\mathbf{0}$ of size $Z_c \times Z_c$; and each element of value 1 in $\mathbf{H_{BG}}$ is replaced by circular permutation matrix $\mathbf{I}(P_{i,j})$ of size $Z_c \times Z_c$, where i and j are the row and column indices of the element, and $\mathbf{I}(P_{i,j})$ is obtained by circularly shifting the identity matrix $\mathbf{I}$ of size $Z_c \times Z_c$ to the right $P_{i,j}$ times. The value of $P_{i,j}$ is given by $P_{i,j} = V_{i,j} \bmod Z_c$ and the value of $V_{i,j}$ is given in [7].
5. Set $d_{k-2Z_c} = w_{k-K}, \forall k = K, \ldots, N + 2Z_c - 1$.

The performance of the NR LDPC codes over an AWGN channel was evaluated using a normalized min-sum decoder, layered scheduling, and a maximum of 20 decoder iterations. Fig. 4.51 shows the required SNR to achieve certain BLER targets as a function of information block size K for code rate 1/2 and QPSK modulation. The results show that the NR LDPC codes provide consistently good performance over a large range of block sizes. For this code rate, base graph 2 is used for all K for which it is defined, that is, for $K \leq 3840$, while base graph 1 is used for larger values of K (note the discontinuity point in the curves at $K = 3840$). As shown in the figure, there is a small gap in performance at the block size where the transition occurs between base graph 1 and base graph 2 [55].

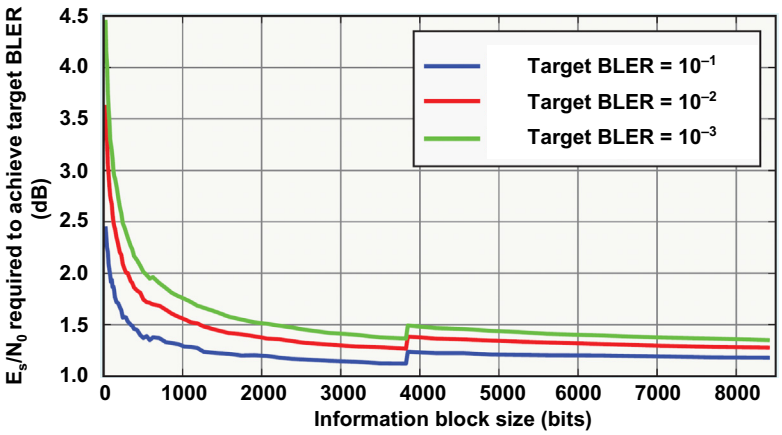


Figure 4.51
Performance of NR LDPC codes at code rate 1/2, QPSK modulation, and 20 iterations [55].

4.1.7.5 Modulation Schemes and MCS Determination

3GPP NR supports various modulation schemes (QPSK, 16QAM, 64QAM, and 256QAM) for CP-OFDM in both downlink and uplink. For the DFT-s-OFDM, however, NR uses an additional modulation $\pi/2$-BPSK in the uplink to achieve better efficiency for power amplifiers and lower PAPR in very low data rate cases. While 1024QAM can theoretically provide 25% throughput gain relative to 256QAM, due to real-world implementation complexity and the need for a very high SINR levels to achieve acceptable BLER targets, it was not included in the NR modulation schemes. A constellation diagram is a graphical representation of the complex envelope of each possible symbol state. The power efficiency is related to the minimum distance between the points in the constellation. The bandwidth efficiency is related to the number of points in the constellation. The gray coding is used to assign groups of bits to each constellation point. In gray coding, adjacent constellation points differ by a single bit. The modulation mapping function takes input bit sequence $b(i)b(i+1)\cdots b(i+Q_m-1)$ and generates the corresponding complex-valued modulation symbol $x=\gamma(I+jQ)$ in the output with the value of γ is chosen to achieve equal average power. More specifically, the NR supports the following modulation schemes depending on the channel conditions experienced by the users [6]:

- *$\pi/2$-BPSK:* In this case, bit $b(i)$ is mapped to complex-valued modulation symbol $d(i)$ according to $d(i)=1/\sqrt{2}[(1-2b(i))+j(1-2b(i))]\exp\left(j\left(\pi/2\right)(i \bmod 2)\right)$.
- *BPSK:* In this case, bit $b(i)$ is mapped to complex-valued modulation symbol $d(i)$ according to $d(i)=1/\sqrt{2}[(1-2b(i))+j(1-2b(i))]$.
- *QPSK:* In this case, a pair of bits $b(2i), b(2i+1)$ is mapped to complex-valued modulation symbol $d(i)$ according to $d(i)=1/\sqrt{2}[(1-2b(2i))+j(1-2b(2i+1))]$.
- *16QAM:* In this case, bit quadruplet $b(4i), b(4i+1), b(4i+2), b(4i+3)$ are mapped to complex-valued modulation symbol $d(i)$ according to $d(i)=1/\sqrt{10}\{(1-2b(4i))[2-(1-2b(4i+2))]+j(1-2b(4i+1))[2-(1-2b(4i+3))]\}$.
- *64QAM:* In this case, bit sextuplet $b(6i), b(6i+1), b(6i+2), b(6i+3), b(6i+4), b(6i+5)$ are mapped to complex-valued modulation symbol $d(i)$ according to $d(i)=1/\sqrt{42}\{(1-2b(6i))[4-(1-2b(6i+2))[2-(1-2b(6i+4))]]+j(1-2b(6i+1))[4-(1-2b(6i+3))[2-(1-2b(6i+5))]]\}$.
- *256QAM:* In this case, bit octuplet $b(8i), b(8i+1), b(8i+2), b(8i+3), b(8i+4), b(8i+5), b(8i+6), b(8i+7)$ are mapped to complex-valued modulation symbol $d(i)$ according to $d(i)=1/\sqrt{170}\{1-2b(8i)[8-1-2b(8i+2)[4-1-2b(8i+4)[2-1-2b(8i+6)]]]+j1-2b(8i+1)[8-1-2b(8i+3)[4-1-2b(8i+5)[2-1-2b(8i+7)]]]\}$.

To determine the modulation order, target code rate, and TBS in the PDSCH, the UE needs to read the 5-bit modulation and coding scheme field I_{MCS} in the DCI to determine the modulation order Q_m and target code rate R based on the procedure that we defined in the

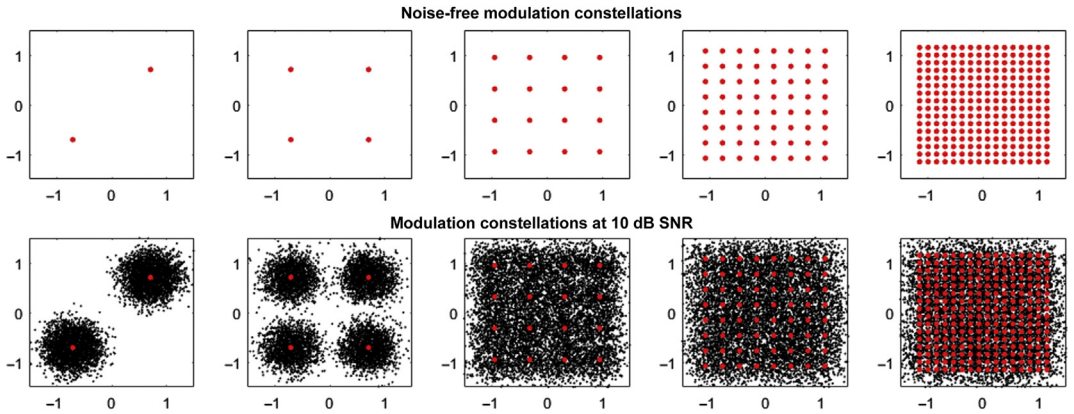


Figure 4.52
Modulation constellations at noise-free and SNR = 10 dB conditions.

previous section. The UE then use the number of layers v and the total number of allocated PRBs before rate matching n_{PRB} to determine to the TBS. The UE may skip decoding of a TB in an initial transmission, if the effective channel code rate is higher than 0.95. The effective channel code rate is defined as the number of downlink information bits (including CRC bits) divided by the number of physical channel bits transmitted on PDSCH [9]. Fig. 4.52 shows the constellations of NR modulation schemes in noise-free and SNR = 10 dB conditions, which demonstrate the effect of achievable SNR at the receiver detector and the choice of modulation order for transmission.

The concepts of MCS, code rate, TB, and TBS in 3GPP NR are similar to those of 3GPP LTE. In NR, the DL-SCH and UL-SCH MCS and code rate for transmission are determined by predefined tables given in [9]. However, TBS determination in NR is more complicated than that of LTE. Unlike LTE where all possible TBSs were precalculated and listed in the MCS table, in NR, the TBS determination process is described as a procedure that has been illustrated in Fig. 4.53. As shown in the flow chart, the initial input to this algorithm is N_{info}; however, to determine this value, the following calculations are necessary [9]:

$$\underbrace{N'_{RE}}_{\substack{\text{The number of REs allocated for}\\ \text{PDSCH within a PRB}}} = 12 \quad \underbrace{N^{symb}_{slot}}_{\substack{\text{The number of symbols of the PDSCH}\\ \text{allocation within the slot}}} - \underbrace{N^{PRB}_{DM-RS}}_{\substack{\text{The number of REs for DM}-\text{RS per PRB}\\ \text{in the scheduled duration including the overhead}\\ \text{of the DM}-\text{RS CDM groups without data}}} - \underbrace{N^{PRB}_{overhead}}_{\substack{\text{The overhead configured by higher}-\text{layer}\\ \text{parameter xOverhead in PDSCH}-\text{ServingCellConfig}}}$$

$$\underbrace{N_{RE}}_{\text{The total number of REs allocated for PDSCH}} = \min(156, N'_{RE}) \quad \underbrace{n_{PRB}}_{\text{The total number of allocated PRBs for the UE}}$$

$$\underbrace{N_{info}}_{\text{Intermediate number of information bits}} = N_{RE} \quad \underbrace{R}_{\text{Target code rate}} \quad \underbrace{Q_m}_{\text{Modulation order}} \quad \underbrace{v}_{\text{Number of layers}}$$

For downlink shared channel, the supported modulation schemes include QPSK, 16QAM, 64QAM, and 256QAM. After detecting the CSI-RS and estimating the channel quality, UE

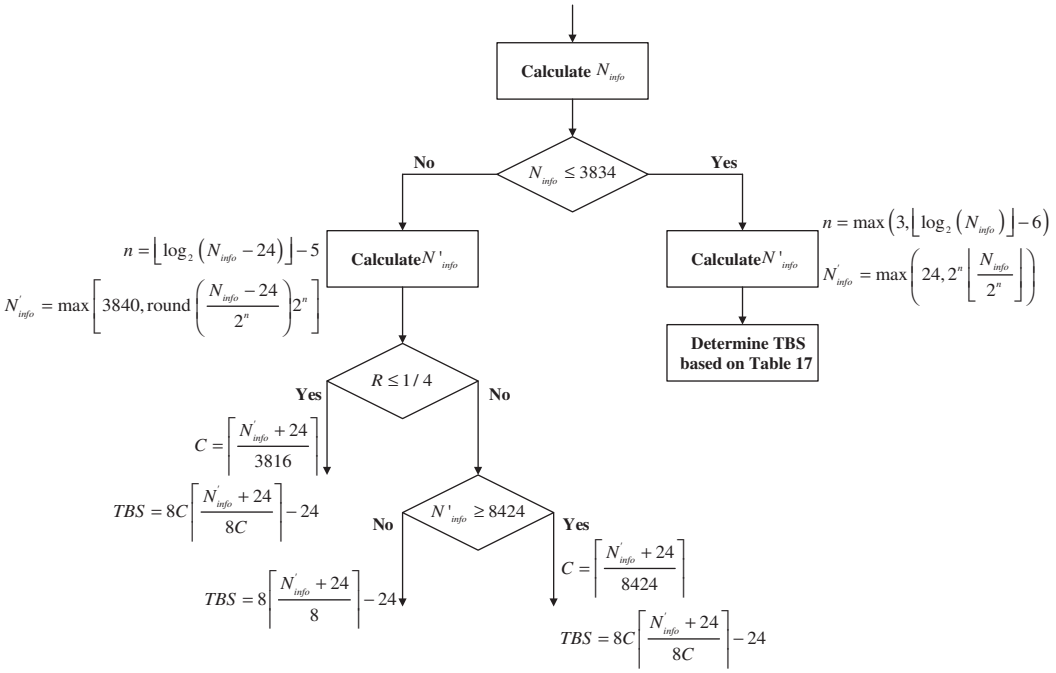


Figure 4.53
TBS determination procedure in NR [30].

reports the CQI to the gNB, which includes the information such as modulation scheme and coding rate. To balance the overhead and the granularity of CQI indication, two CQI/MCS tables are defined for eMBB, where the maximum order of modulation in one CQI/MCS table is 64QAM and in another table is 256QAM (see MCS Tables I and II in Table 4.16). The network will instruct the UE to select CQI/MCS table through RRC signaling. The third MCS table is meant for URLLC use cases where the target BLER is 10^{-5}, which is signaled to UE when the CRC of the PDCCH is scrambled with MCS-C-RNTI. This MCS table was designed to allow single transmissions to the UEs with delay sensitive applications to ensure maximum likelihood of correct reception.

Given the modulation order, the number of resource blocks scheduled, and the scheduled transmission duration, the number of available resource elements can be computed. From this number, the resource elements used for DM-RS are subtracted. A constant, configured by higher layers, modeling the overhead of other signals such as CSI-RS or SRS, is also subtracted. The resulting estimate of resource elements available for data allocation is then, together with the number of transmission layers, the modulation order, and the code rate obtained from the MCS table, are used to calculate an intermediate number of information bits. This intermediate number is then quantized to obtain the final transport block size,

Table 4.16: MCS index tables 1/2/3 for physical downlink shared channel [9].

MCS Index I_{MCS}	Modulation Order Q_m	Target Code Rate $1024 \times R$	Spectral Efficiency	Modulation Order Q_m	Target Code Rate $1024 \times R$	Spectral Efficiency	Modulation Order Q_m	Target Code Rate $1024 \times R$	Spectral Efficiency
	MCS Table I			MCS Table II			MCS Table III		
0	2	120	0.2344	2	120	0.2344	2	30	0.0586
1	2	157	0.3066	2	193	0.3770	2	40	0.0781
2	2	193	0.3770	2	308	0.6016	2	50	0.0977
3	2	251	0.4902	2	449	0.8770	2	64	0.1250
4	2	308	0.6016	2	602	1.1758	2	78	0.1523
5	2	379	0.7402	4	378	1.4766	2	99	0.1934
6	2	449	0.8770	4	434	1.6953	2	120	0.2344
7	2	526	1.0273	4	490	1.9141	2	157	0.3066
8	2	602	1.1758	4	553	2.1602	2	193	0.3770
9	2	679	1.3262	4	616	2.4063	2	251	0.4902
10	4	340	1.3281	4	658	2.5703	2	308	0.6016
11	4	378	1.4766	6	466	2.7305	2	379	0.7402
12	4	434	1.6953	6	517	3.0293	2	449	0.8770
13	4	490	1.9141	6	567	3.3223	2	526	1.0273
14	4	553	2.1602	6	616	3.6094	2	602	1.1758
15	4	616	2.4063	6	666	3.9023	4	340	1.3281
16	4	658	2.5703	6	719	4.2129	4	378	1.4766
17	6	438	2.5664	6	772	4.5234	4	434	1.6953
18	6	466	2.7305	6	822	4.8164	4	490	1.9141
19	6	517	3.0293	6	873	5.1152	4	553	2.1602
20	6	567	3.3223	8	682.5	5.3320	4	616	2.4063
21	6	616	3.6094	8	711	5.5547	6	438	2.5664
22	6	666	3.9023	8	754	5.8906	6	466	2.7305
23	6	719	4.2129	8	797	6.2266	6	517	3.0293
24	6	772	4.5234	8	841	6.5703	6	567	3.3223
25	6	822	4.8164	8	885	6.9141	6	616	3.6094
26	6	873	5.1152	8	916.5	7.1602	6	666	3.9023
27	6	910	5.3320	8	948	7.4063	6	719	4.2129
28	6	948	5.5547	2	Reserved		6	772	4.5234
29	2	Reserved		4	Reserved		2	Reserved	
30	4	Reserved		6	Reserved		4	Reserved	
31	6	Reserved		8	Reserved		6	Reserved	

Table 4.17: NR transport-block size for $N_{info} \leq 3824$ [9].

Index	TBS (bits)	Index	TBS (bits)	Index	TBS (bits)	Index	TBS (bits)	Index	TBS (bits)	Index	TBS (bits)	Index	TBS (bits)
1	24	16	144	31	336	46	704	61	1288	76	2216	91	3624
2	32	17	152	32	352	47	736	62	1320	77	2280	92	3752
3	40	18	160	33	368	48	768	63	1352	78	2408	93	3824
4	48	19	168	34	384	49	808	64	1416	79	2472		
5	56	20	176	35	408	50	848	65	1480	80	2536		
6	64	21	184	36	432	51	888	66	1544	81	2600		
7	72	22	192	37	456	52	928	67	1608	82	2664		
8	80	23	208	38	480	53	984	68	1672	83	2728		
9	88	24	224	39	504	54	1032	69	1736	84	2792		
10	96	25	240	40	528	55	1064	70	1800	85	2856		
11	104	26	256	41	552	56	1128	71	1864	86	2976		
12	112	27	272	42	576	57	1160	72	1928	87	3104		
13	120	28	288	43	608	58	1192	73	2024	88	3240		
14	128	29	304	44	640	59	1224	74	2088	89	3368		
15	136	30	320	45	672	60	1256	75	2152	90	3496		

while at the same time ensuring byte-aligned code blocks, and that no padding bits are needed in the LDPC coding. The quantization also results in the same transport block size being obtained, even if there are small variations in the amount of resources allocated, a property that is useful when scheduling retransmissions on a different set of resources than the initial transmission (see Table 4.17). In the case of a retransmission, the transport block size by definition, is unchanged and there is no need to signal this information. Instead, the reserved entries represent the modulation scheme (QPSK, 16QAM, 64QAM, or if configured 256QAM), which allows the scheduler to use an (almost) arbitrary combination of resource blocks for the retransmission. The use of the reserved entries assumes that the UE properly received the control signaling for the initial transmission, if this is not the case, the retransmission should explicitly indicate the transport block size [14].

4.1.8 HARQ Operation and Protocols

4.1.8.1 HARQ Principles

While ARQ error control mechanism is simple and provides high transmission reliability, the throughput of ARQ schemes drop rapidly with increasing channel error rates, and the latency, due to retransmissions, could be excessively high and intolerable for some delay--sensitive applications. Systems using forward error correction (FEC), on the other hand, can maintain constant throughput regardless of channel error rate. However, FEC schemes have some drawbacks. High reliability is hard to achieve with FEC and requires the use of long and powerful error correction codes that increase the complexity of implementation.

The drawbacks of ARQ and FEC can be overcome, if the two error control schemes are properly combined. In order to achieve increased throughput and lower latency in packet transmission, hybrid ARQ (HARQ) scheme was designed to combine ARQ error-control mechanism and FEC coding. A HARQ system consists of a FEC subsystem contained in an ARQ system. In this approach, the average number of retransmissions is reduced by using FEC through correction of the error patterns that occur more frequently; however, when the less frequent error patterns are detected, the receiver requests a retransmission where each retransmission carries the same or some redundant information to help the packet detection. The HARQ uses FEC to correct a subset of errors at the receiver and rely on error detection to detect the remaining errors. Most practical HARQ schemes utilize CRC codes for error detection and some form of FEC for correcting the transmission errors. The HARQ schemes are typically classified into two groups depending on the content of subsequent retransmissions, as follows [15]:

1. *HARQ with chase combining:* In this HARQ scheme, the same data packet is transmitted in all retransmissions. Soft combining may be used to improve the reliability. The blocks of data along with the CRC code are encoded using FEC encoder before transmission. If the receiver is unable to correctly decode the data block, a retransmission is requested. When a retransmitted coded block is received, it is combined with the previously received block corresponding to the same information bits (using for example maximum ratio combining method) and fed to the decoder. Since each retransmission is an identical replica of the original transmission, the received E_b/N_0, that is, the energy per information bit divided by the noise spectral power density, increases per each retransmission, improving the likelihood of correct decoding. In chase combining HARQ, the redundancy version of the encoded bits is not changed from one transmission to the next; therefore, the puncturing pattern remains the same. The receiver uses the current and all previous HARQ transmissions of the code block in order to decode the information bits. The process continues until either the information bits are correctly decoded and pass the CRC test or the maximum number of HARQ retransmissions is reached. When the maximum number of retransmissions is reached, the MAC sublayer resets the process and continues with fresh transmission of the same code block. A number of parallel channels for HARQ can help improve the throughput as one process is awaiting an acknowledgment; another process can utilize the channel and transmit subpackets. Fig. 4.54 illustrates the operation of chase combining HARQ (HARQ-CC) scheme and how the retransmission of the same coded bits changes the combined energy per bit E_b while maintaining the effective code rate intact.
2. *HARQ with incremental redundancy:* In this HARQ scheme, additional parity bits are sent in subsequent retransmissions. Therefore, after each retransmission, a richer set of parity bits is available at the receiver, improving the probability of reliable decoding. In incremental redundancy schemes, however, information cannot be recovered from parity

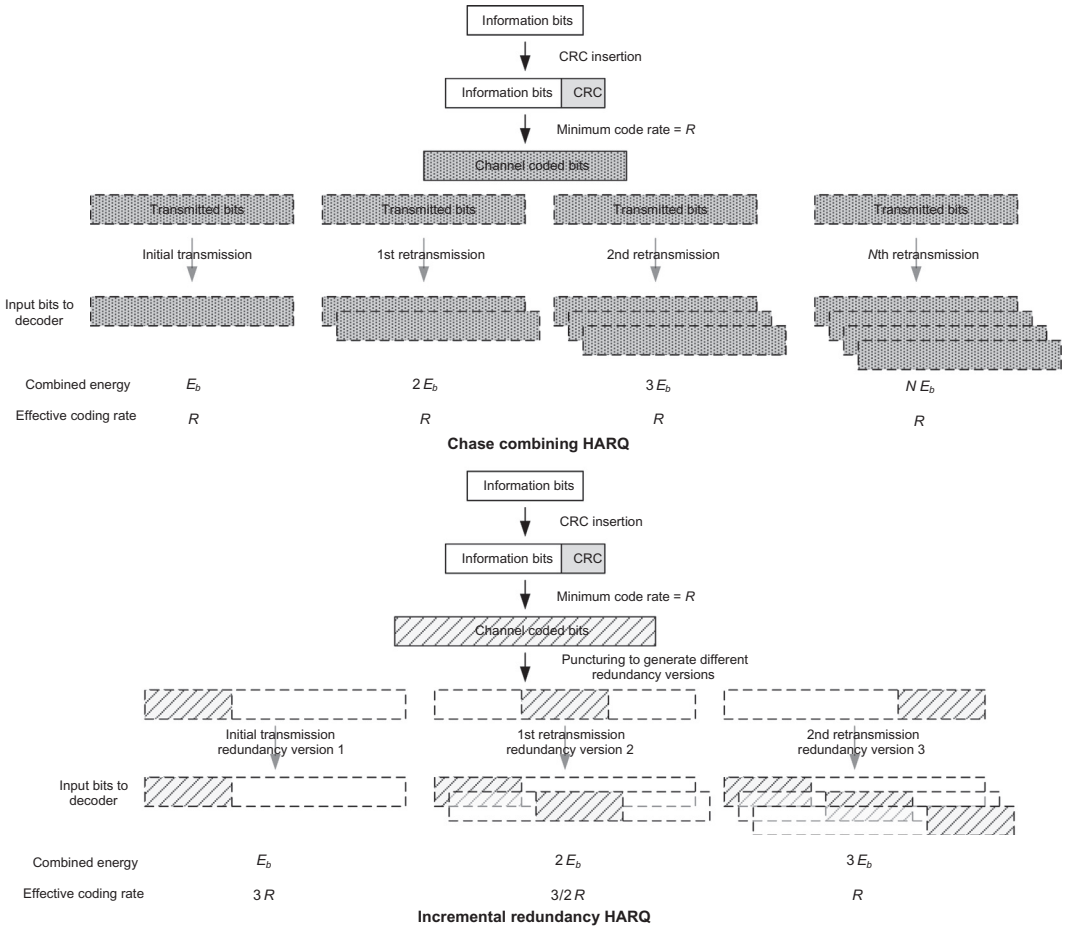


Figure 4.54
Illustration of the chase combining and incremental redundancy HARQ schemes [15].

bits alone. In incremental redundancy HARQ (HARQ-IR) scheme, a number of coded bits with increasing redundancy are generated and transmitted to the receiver when a retransmission is requested to assist the receiver with the decoding of the information bits. The receiver combines each retransmission with the previously received bits belonging to the same packet. Since each retransmission carries additional parity bits, the effective code rate is lowered by each retransmission as shown in Fig. 4.54. The IR is based on low-rate code and the different redundancy versions are generated by puncturing the channel coder output. In the example shown in Fig. 4.54, the basic code rate is R, and one-third of the coded bits are transmitted in each retransmission. Aside from increasing the received signal to noise ratio E_b/N_0 by each retransmission due to

combining, there is a coding gain[13] attained as a result of each retransmission. It must be noted that chase combining is a special case of HARQ-IR where the retransmissions are identical copies of the original coded bits.

4.1.8.2 UE Processing Times, HARQ Protocol and Timing

3GPP NR uses an asynchronous HARQ-IR scheme in the downlink and uplink. The gNB provides the UE with the HARQ-ACK feedback timing either dynamically in the DCI or semi-statically through RRC configuration messages. The gNB schedules each uplink transmission and retransmission using the uplink grant on DCI. In LTE, the basic mode of operation for uplink HARQ is synchronous retransmission, which can be used to reduce the scheduling overhead for retransmissions. In this case, HARQ ACK/NACK is carried on PHICH as a short and efficient message. In NR, asynchronous HARQ is supported. In order to support asynchronous HARQ, a straightforward solution for the gNB is to send an explicit uplink grant through PDCCH for the retransmission in the same way that is done for transmissions in LTE. In some sense, the explicit grant can imply an implicit ACK/NACK. For example, an explicit scheduling grant of a retransmission may imply a NACK for the initial transmission. The maximum number of HARQ processes in the downlink and uplink per cell is 16. The number of HARQ processes is separately configured for the UE for each cell by RRC parameter *nrofHARQ-processesForPDSCH*. In the absence of any configuration, the UE may assume a default number of 8 HARQ processes.

The UE must provide a valid HARQ-ACK message, if the first uplink symbol of PUCCH conveying the HARQ-ACK information, as identified by HARQ-ACK timing parameter K_1 and the assigned PUCCH resource including the effect of the timing advance, starts on or after symbol L_1, that is, the next uplink symbol with its cyclic prefix starting after $T_{proc} = 2192(N_1 + d_x)\kappa 2^{-\mu}T_c$ (processing time) following the end of the last symbol of the PDSCH carrying the transport block being acknowledged. As shown in Fig. 4.55, parameter N_1 is based on the numerology and corresponds to $(\mu_{PDCCH}, \mu_{PDSCH}, \mu_{UL})$ where μ_{PDCCH}, μ_{PDSCH}, and μ_{UL} correspond to the subcarrier spacing of the PDCCH scheduling, PDSCH transmission, and the uplink channel on which the HARQ-ACK is transmitted, respectively. As shown in Table 4.18, the value of parameter N_1 further depends on the PDSCH DM-RS pattern and whether additional DM-RS is used as well as UE PDSCH processing capability. The value of parameter d_x is dependent on the PDSCH mapping type (A or B), UE PDSCH processing capability, and the number of PDSCH symbols [9]. The timing relationship between HARQ-ACK and PDSCH data transmission depends on the value of the above parameters and is depicted in Fig. 4.55.

[13] The coding gain is defined as the difference between E_b/N_0 required to achieve a given bit error rate in a coded system and the E_b/N_0 required to achieve the same BER in an uncoded system.

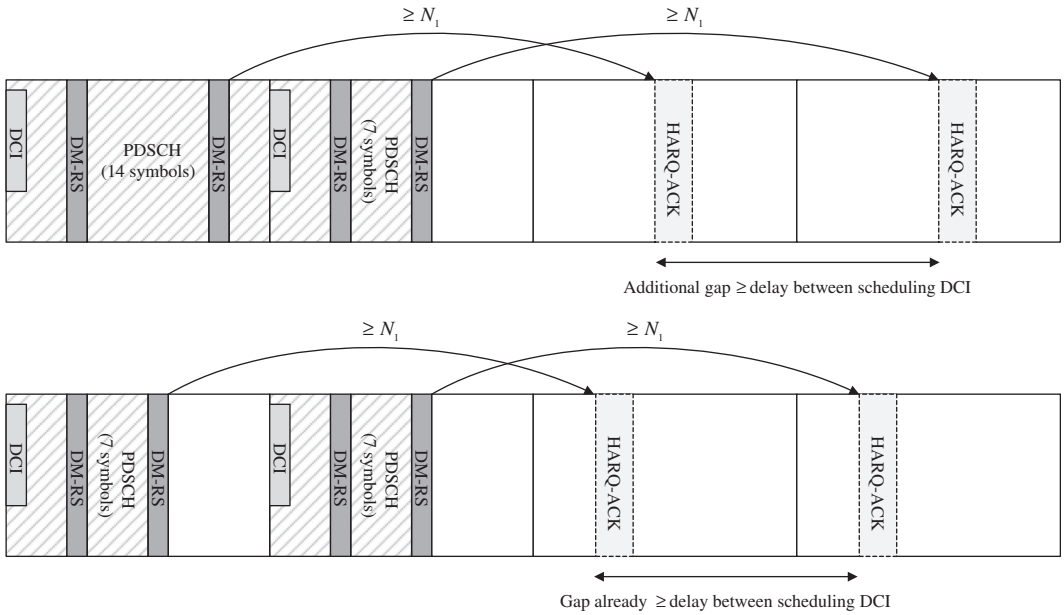


Figure 4.55
Timing relationship between PDSCH and HARQ-ACK transmission [9].

Table 4.18: Physical downlink shared channel processing times [9].

μ	PDSCH Decoding Time N_1 (OFDM Symbols)		PDSCH Decoding Time N_1 (OFDM Symbols)
	PDSCH Processing Capability 1		PDSCH Processing Capability 2
	No Additional PDSCH	Additional PDSCH	No Additional PDSCH
	DM-RS Configured (Front-Loaded DM-RS)	DM-RS Configured	DM-RS Configured (Front-Loaded DM-RS)
0	8	13	3
1	10	13	4.5
2	17	20	9 (in Frequency Range 1)
3	20	24	–

In the case of carrier aggregation, the multi-carrier nature of the physical layer is only exposed to the MAC sublayer for which one HARQ entity is required per serving cell. In both uplink and downlink, there is one independent HARQ entity per serving cell and one TB is generated per TTI in the absence of spatial multiplexing. Each TB and its associated HARQ retransmissions are mapped to a single serving cell [11]. Fig. 4.56 depicts HARQ-ACK timing requirements in a cross-carrier scheduling scenario when a UE receives and transmits to from/to two gNBs.

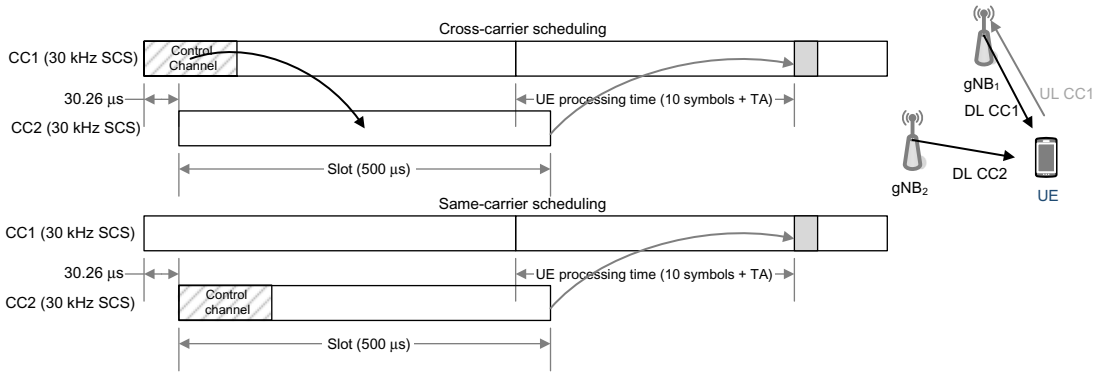


Figure 4.56
UE processing time considering timing difference between different cells [65].

In the NR downlink, retransmissions are scheduled in the same way as new data, that is, they may occur at any time and at an arbitrary frequency location within the downlink cell bandwidth. The scheduling assignment contains the necessary HARQ-related control signaling such as process number, new-data indicator, CBGTI, and CBGFI in case of CBG-based retransmission as well as information to handle the transmission of the acknowledgment in the uplink such as timing and resource indication. Upon receiving a scheduling assignment in the DCI, the receiver would attempt to decode the TB after soft combining with previous retransmissions. Since transmissions and retransmissions are scheduled using the same framework, the UE needs to know whether the transmission is a new transmission, in which case the soft buffer should be flushed, or a retransmission, in which case soft combining should be performed. Therefore, an explicit new-data indicator is included for the scheduled TB as part of the scheduling information transmitted in the downlink. The new-data indicator is toggled for a new TB. Upon reception of a downlink scheduling assignment, the UE checks the new-data indicator to determine whether the current transmission should be soft combined with the received data currently in the soft buffer for the HARQ process or the soft buffer should be cleared [9,11].

The NR uplink uses asynchronous HARQ protocol in the same way as downlink. The HARQ-related information including process number, new-data indicator, CBG-based retransmission (if configured), and the CBGTI are included in the scheduling grant. The uplink CBGTI is used in the same way as in the downlink, that is, to indicate the CBGs that need to be retransmitted in the case of CBG-based retransmission. Note that no CBGFI is needed in the uplink as the soft buffer is located in the gNB which can decide whether to flush the buffer or not based on the scheduling decisions.

The use of HARQ in NR allows reliable delivery of layer-1 packets between peer entities. Each HARQ process supports one TB when the physical layer is not configured for

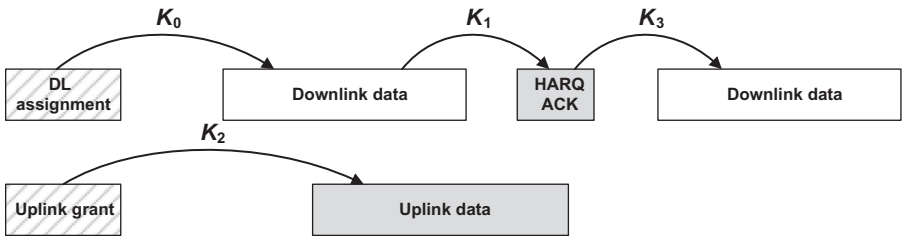


Figure 4.57
NR HARQ protocol timing [9].

downlink/uplink spatial multiplexing; otherwise, each HARQ process supports one or multiple TBs. The NR HARQ operation and timing are illustrated in Fig. 4.57, where the parameters used in the figure are defined as follows [8,9]:

- K_0 denotes the delay between downlink grant and corresponding PDSCH data reception.
- K_1 is the delay between PDSCH data reception and the corresponding ACK/NACK transmission on the uplink.
- K_2 denotes the delay between uplink grant reception in downlink and the uplink data transmission on PUSCH.
- K_3 is the delay between ACK/NACK reception in the uplink and the corresponding retransmission of downlink data on PDSCH.
- The parameters K_0, K_1, and K_2 are signaled via DCI and if $K_1 = 0$, a self-contained subframe/slot is configured.

The choice of N_1 value has a significant impact on the UE processing time, where N_1 is defined as the number of OFDM symbols from the end of PDSCH reception to the start of the corresponding ACK/NACK transmission from UE (as shown Fig. 4.58). Depending on the frame structure, some UE data processing can be done in parallel with PDSCH reception in order to allow faster HARQ ACK/NACK transmission. For example, for front-loaded DM-RS pattern and slot-based scheduling with subcarrier spacing of 15 kHz, it can be shown that PDCCH processing, the demodulation/detection of the symbols other than the last symbol of PDSCH can be performed in $T_1 = T_{FFT} + T_{demodulation} + T_{decode} + T_{UL\text{-}HARQ} + T_{other}$ where T_{FFT}, $T_{demodulation}$, T_{decode}, $T_{UL\text{-}HARQ}$, and T_{other} denote the processing time for FFT/IFFT per symbol, the demodulation time for one symbol, the decoding time for one symbol code blocks, the processing time for uplink ACK/NACK, and the other implementation-specific processing times, respectively.

For the non-slot-based scheduling (e.g., two-symbol mini-slot) and subcarrier spacing 15 kHz, it can be shown that the UE processing requirement $T_2 = T_{PDCCH} + 2T_{demodulation} + T_{decode} + T_{UL\text{-}HARQ} + T_{other}$ where T_{PDCCH} denotes the

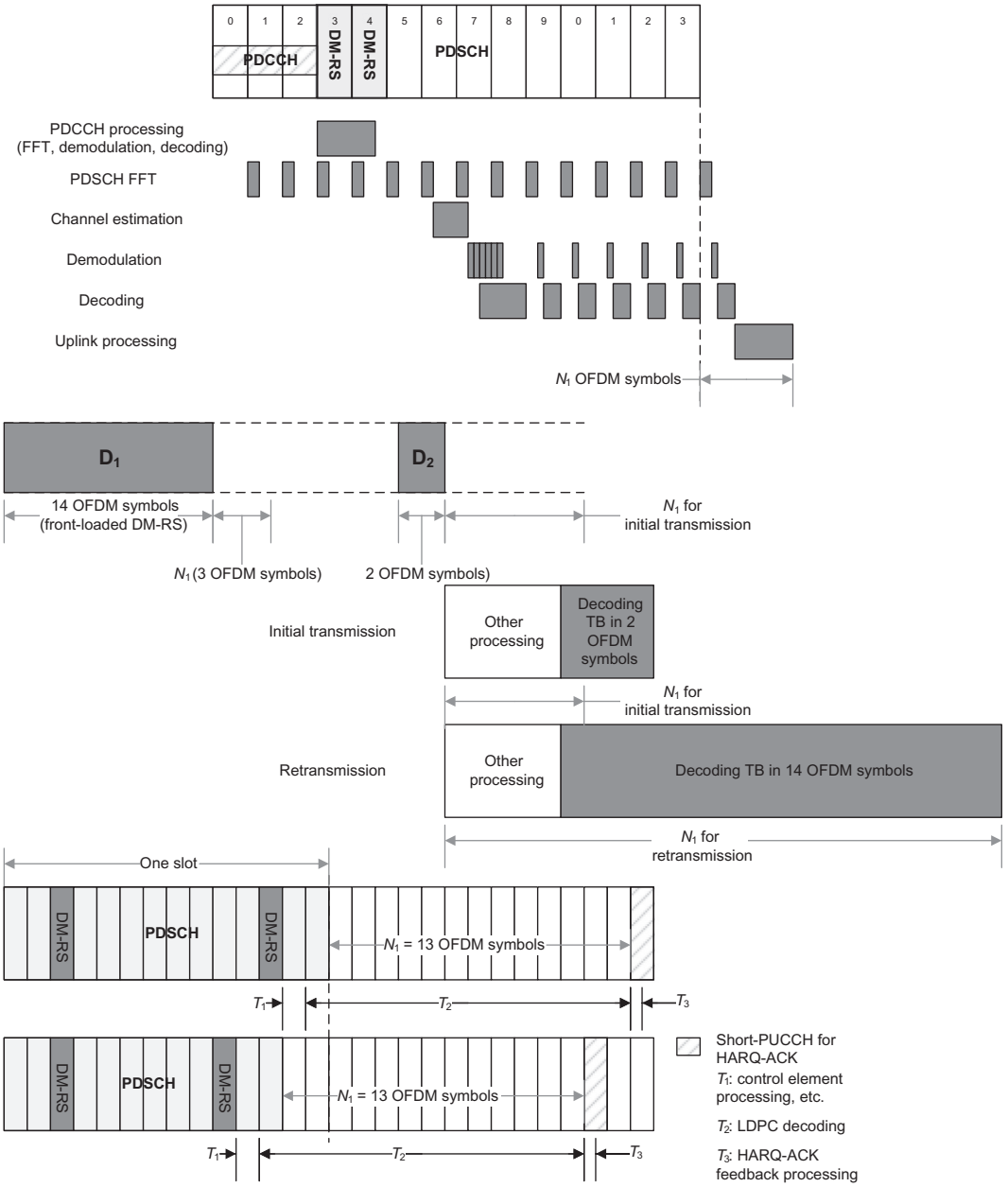


Figure 4.58
UE processing time components (front-loaded DM-RS, slot-based scheduling, and 15 kHz SCS).

processing time for PDCCH including decoding, demodulation, and parsing. The above processing time calculation was done under certain conditions which may vary in different scenarios. There is no limit for scheduling transmission/retransmission; however, the initial transmission is slot-based, and the retransmissions may be non-slot based as shown in Fig. 4.58. In the following example, we assume subcarrier spacing is 15 kHz, D_1 is a slot-based downlink scheduling period with front-loaded DM-RS and D_2 is a non-slot-based scheduling period. If D_2 is the initial transmission then the processing time is calculated as T_2; nevertheless, if D_2 is the retransmission of D_1 then we need to perform TB decoding with 14 OFDM symbols; thus the processing time for retransmission is shown to be $T_3 = T_{PDCCH} + 2T_{demodulation} + 13T_{decode} + T_{UL\text{-}HARQ} + T_{other}$, meaning that for the same conditions, the processing time for initial transmission and retransmission are different [65].

As we mentioned earlier, the maximum number of HARQ processes per carrier supported in NR is 8 or 16. For continuous downlink transmission at the peak data rate, the minimum number of HARQ processes is $\min(N_{DL-HARQ}) = K_1 + K_3 + \lfloor 2T_d/TTI_{DL} \rfloor$, in which T_d denotes the transmission delay. The required number of HARQ processes may vary depending on UE HARQ processing capability, numerology, and network configurations. The determination of the number of HARQ processes is up to gNB scheduler and thus signaled via the DCI. To reduce the overhead, the gNB can semi-statically configure a UE with a smaller number of HARQ processes than 16 per bandwidth part. In order to reduce the latency due to HARQ retransmissions and to avoid retransmission of the entire TB and the performance degradation of HARQ due to large transport block size, NR defined CBG-based transmission and HARQ operation which supports single-/multi-bit HARQ-ACK feedback in Rel-15. The CBG-based (re)transmissions are only allowed for the same TB of a HARQ process. The CBG can include all code blocks of a TB regardless of the size of the TB. In such conditions, the UE reports single HARQ-ACK bits for the TB. The CBG can include one code block and its granularity is configurable. The UE is semi-statically configured by RRC signaling to enable CBG-based retransmission.

When the *CSI request* field in a DCI triggers a CSI report(s) on PUSCH, the UE is required to provide a valid CSI report(s), if the first uplink symbol to carry the corresponding CSI report(s), including the effect of the timing advance, starts no earlier than at symbol l_{CSI}, and if the first uplink symbol to carry the corresponding CSI report, including the effect of the timing advance, starts no earlier than at symbol l'_{CSI}. The reference symbol l_{CSI} is defined as the next uplink symbol with starting $T_{proc\text{-}CSI} = 2192 l_{CSI} \kappa 2^{-\mu} T_C$ after the end of the last symbol of the PDCCH triggering the CSI report. The reference symbol l'_{CSI} is defined as the next uplink symbol starting $T'_{proc\text{-}CSI} = 2192 l'_{CSI} \kappa 2^{-\mu} T_C$ after the end of the last symbol of the latest aperiodic CSI-RS resource for channel measurements, aperiodic CSI-IM used for interference measurements, and aperiodic NZP CSI-RS for interference measurement, when aperiodic CSI-RS is used for channel measurement for the triggered nth CSI report [9].

Table 4.19: NR channel state information computation delay requirements 1 and 2 [9].

μ	Delay Requirement 1		Delay Requirement 2			
	Z_1 (Symbols)		Z_1 (Symbols)		Z_2 (Symbols)	
	Z_1	Z'_1	Z_1	Z'_1	Z_2	Z'_2
0	10	8	22	16	40	37
1	13	11	33	30	72	69
2	25	21	44	42	141	140
3	43	36	97	85	152	140

As a result, l_{CSI} and l'_{CSI} are defined as $l_{CSI} = \max_{m=0,1,\ldots,N_{CSI}-1} l_{CSI}(m)$ and $l'_{CSI} = \max_{m=0,1,\ldots,N_{CSI}-1} l'_{CSI}(m)$, where N_{CSI} is the number of updated CSI report(s), $l_{CSI}(m)$ and $l'_{CSI}(m)$ corresponds to the mth updated CSI report. The values of $l_{CSI}(m)$ and $l'_{CSI}(m)$ are set to l_1, l'_1, l_2, or l'_2 depending on the CSI computation delay requirements as shown in Table 4.19, μ corresponds to the $\min(\mu_{PDCCH}, \mu_{CSI\text{-}RS}, \mu_{UL})$ where the μ_{PDCCH} corresponds to the subcarrier spacing of the PDCCH in which the DCI was transmitted, and μ_{UL} corresponds to the subcarrier spacing of the PUSCH in which the CSI report is transmitted, and $\mu_{CSI\text{-}RS}$ corresponds to the minimum subcarrier spacing of the aperiodic CSI-RS triggered by the DCI.

4.1.8.3 Semi-static/Dynamic Codebook HARQ-ACK Multiplexing

The NR supports multiplexing of HARQ acknowledgments for multiple transport blocks into an acknowledgment bitmap, when multiple TBs need to be acknowledged at the same time or alternatively multiple acknowledgments need to be transmitted in the uplink at the same time in carrier aggregation and CBG-based retransmission scenarios. This bitmap can be signaled either via a semi-static codebook or a dynamic codebook both configured through RRC signaling. The semi-static codebook can be viewed as a matrix consisting of a time-domain dimension and a component-carrier, CBG, or MIMO layer dimension, both of which are semi-statically configured. The size in the time domain is given by the maximum and minimum HARQ acknowledgment timings and the size in the carrier domain is given by the number of simultaneous transport blocks or CBGs across all component carriers. An example is shown in Fig. 4.59, where the acknowledgment timings are one, two, three, and four slots, respectively, and three carriers, one with two TBs, one with one TB, and one with four CBGs, are configured. Since the codebook size is fixed, the number of bits to transmit in a HARQ-ACK is known and an appropriate format for the uplink control signaling can be selected. Each entry in the matrix represents successful/unsuccessful outcome of the decoding of the corresponding downlink transmission. A NACK is sent in position of unused transmission opportunities in the codebook, resulting in improved robustness in the case of missed downlink assignment where the gNB can retransmit the missing TB or the CBG [14].

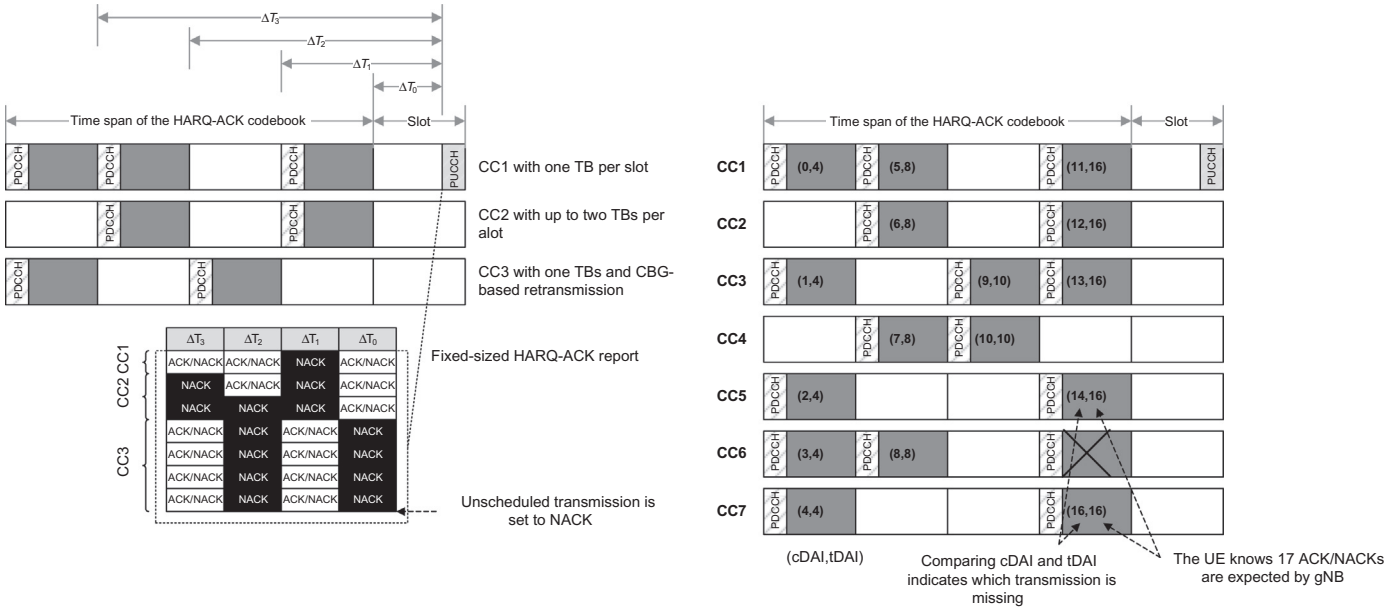


Figure 4.59
Illustration of the semi-static and dynamic codebooks (example) [14].

One drawback of the semi-static codebook is the potentially large size of the HARQ feedback. For a small number of component carriers and no CBG-based retransmissions, this may not be a problem; however, if a large number of carriers or CBGs are configured and only a fraction of them are used, the semi-static codebook will become inefficient. To address the drawback of a potentially large semi-static codebook size in some scenarios, NR also supports a dynamic HARQ-ACK codebook, which is the default HARQ-ACK codebook, unless the system is differently configured. With a dynamic codebook, only the acknowledgment information for the scheduled carriers is included in the HARQ feedback, as opposed to all carriers as is the case for a semi-static codebook. Hence, the size of the codebook may dynamically vary as a function of the number of the scheduled carriers. This would reduce the size of the HARQ acknowledgment message. A dynamic HARQ-ACK codebook would be straightforward option, if there were no errors in the downlink control signaling. However, in the presence of an error in the downlink control signaling, the UE and the gNB may have different understanding of the number of scheduled carriers, which would lead to an incorrect codebook size and possibly corrupted HARQ feedback for all carriers. As an example, assume that a device is scheduled for downlink transmission in two consecutive slots, but the PDCCH of the first slot is not received; thus the scheduling assignment for the first slot is missing. In this case, the UE will transmit an acknowledgment for the second slot only, while the gNB expects to receive acknowledgments for two slots. To mitigate such cases, NR supports DAI which is included in the DCI containing the downlink assignment. The DAI field is further divided into two parts, a counter DAI (cDAI) and, in the case of carrier aggregation, a total DAI (tDAI). The cDAI included in the DCI indicates the number of scheduled downlink transmissions up to the point that the DCI was received in a carrier first, time second order. The tDAI included in the DCI indicates the total number of downlink transmissions across all carriers up to this point in time, that is, the highest cDAI at the current point in time. The cDAI and tDAI are represented with decimal numbers with no limitation. In practice, 2 bits are used for each and the numbering is calculated in modulo four. This can be compared with the semi-static codebook which would require a fixed number of entries regardless of the number of active transmissions. If one transmission on a component carrier is lost, without the DAI mechanism, this would result in mismatched codebooks between the UE and the gNB; however, as long as the device receives at least one component carrier, it knows the value of the tDAI, hence the size of the codebook at this point in time. Furthermore, by checking the values received for the cDAI, it can conclude which component carrier was missed and that a negative acknowledgment should be assumed in the codebook for this position [8].

There are two different types of codebook determination algorithms referred to as Type 1 and Type 2. Each of these types is divided into two cases depending on whether the HARQ-ACK is reported on PUCCH or PUSCH, whose usage are configured and signaled through RRC parameters. The codebook determination algorithm types and the associated RRC parameters are summarized in Table 4.20.

Table 4.20: HARQ-ACK codebook determination [8].

Codebook Determination Type	Condition
CBG-based HARQ-ACK codebook determination	*PDSCH-CodeBlockGroupTransmission* = ON
Type-1 HARQ-ACK codebook determination	*pdsch-HARQ-ACK-Codebook* = *semistatic*
Type-1 HARQ-ACK codebook in PUCCH	–
Type-1 HARQ-ACK codebook in PUSCH	–
Type-2 HARQ-ACK codebook determination	*pdsch-HARQ-ACK-Codebook* = *dynamic*
Type-2 HARQ-ACK codebook in PUCCH	*PDSCH-CodeBlockGroupTransmission* = OFF
Type-2 HARQ-ACK codebook in PUSCH	–

4.1.9 Downlink MIMO Schemes

The new radio downlink control and traffic channels rely on DM-RSs to facilitate coherent detection where a UE can assume that the DM-RSs are jointly precoded with the data. In NR Rel-15, downlink multi-antenna precoding is transparent to the UE and the network can apply any transmitter-side precoding without informing the device about the selected precoder. The specification impact of downlink multi-antenna precoding is therefore mainly related to the measurements and reporting mechanisms conducted by the device to support network selection of precoder for downlink transmissions. These precoder-related measurements and reporting are part of the more general CSI reporting framework based on report configurations which was described in previous sections.

A CSI report may consist of rank indicator (RI), precoder-matrix indicator (PMI), and CQI indicating a suitable transmission rank, precoding matrix given the selected rank, coding and modulation scheme given the selected precoder matrix, respectively, from the device perspective. As we mentioned earlier, the reported PMI is a suitable precoder matrix from the UE perspective to be used for downlink transmissions. Each PMI value corresponds to a specific precoder matrix or codebook. Note that the device selects PMI based on a certain number of antenna ports, given by the number of antenna ports of the configured CSI-RS associated with the report configuration, and the selected rank. There is at least one codebook for each valid combination of antenna ports and rank. It must be understood that the suggested codebooks by the UE do not compel the gNB, in practice, not to use another precoding matrix for downlink transmission to the device.

The MU-MIMO schemes typically require detailed knowledge of the channel experienced by each device at the gNB compared to SU-MIMO precoding and transmission to a single device. Therefore, NR defines two types of CSI that differ in the structure and size of the precoding matrices (codebooks), that is, Type I CSI and Type II CSI. Type I CSI primarily targets scenarios where a single user is scheduled within a given time/frequency resources potentially with transmission of a relatively large number of layers in parallel (high-order spatial multiplexing) and Type II CSI mainly targets MU-MIMO scenarios with multiple

devices being scheduled simultaneously within the same time/frequency resources but with only a limited number of spatial layers (maximum of two layers) per scheduled device.

The codebooks for Type I CSI are relatively simple and used to focus the transmit energy at the target receiver. Inter-layer interference is assumed to be primarily handled by utilizing multiple receive antennas and advanced receiver architectures. In contrast, the codebooks for Type II CSI are significantly more extensive, allowing the PMI to provide channel information with much higher spatial granularity. The more extensive channel information allows the network to select a downlink precoder that not only focuses the transmitted energy at the target device but also limits the interference to other devices simultaneously scheduled on the same time/frequency resources. The higher spatial granularity of the PMI feedback is made possible with significantly higher signaling overhead. While a PMI report for Type I CSI may consist of a few tens of bits, the PMI report for Type II CSI may consist of several hundred bits. Therefore, Type II CSI is mainly applicable for low-mobility scenarios where the feedback periodicity in time can be reduced.

4.1.9.1 Capacity of MIMO Channels

A generic MIMO system consists of a MIMO transmitter with N_{TX} transmit antennas, a MIMO receiver with N_{RX} receive antennas, and $N_{RX} \times N_{TX}$ paths or channels between transmit and receive antennas. Let $x_k(t)$ denote the transmitted signal from the kth transmit antenna at time t then the received signal at the lth antenna can be expressed as $y_l(t) = \sum_{k=0}^{N_{TX}-1} h_{lk}(t)^* x_k(t) + n_l(t)$ where $h_{lk}(t)$ and $n_l(t)$ are the channel impulse response between the kth transmit antenna and lth receive antenna and the additive noise at the lth receive antenna port, respectively. The preceding equation can be written in the frequency-domain as $Y_l(\omega) = H_{lk}(\omega)X_k(\omega) + N_l(\omega)$. If $\mathbf{x}(\omega) = \left[X_1(\omega), X_2(\omega), \ldots, X_{N_{TX}}(\omega)\right]^T$, $\mathbf{y}(\omega) = \left[Y_1(\omega), Y_2(\omega), \ldots, Y_{N_{RX}}(\omega)\right]^T$, and $\mathbf{n}(\omega) = \left[N_1(\omega), N_2(\omega), \ldots, N_{N_{RX}}(\omega)\right]^T$ denote the Fourier transform vectors of $\mathbf{x}_k(t)$, $\mathbf{y}_l(t)$, and $\mathbf{n}_l(t)$, respectively then $\mathbf{y}(\omega) = \mathbf{H}(\omega)\mathbf{x}(\omega) + \mathbf{n}(\omega)$ where $\mathbf{H}(\omega)$ is an $N_{RX} \times N_{TX}$ channel matrix with $H_{lk}(\omega)|_{k=1,2,\ldots,N_{TX};l=1,2,\ldots,N_{RX}}$ entries. Assuming a linear time-invariant MIMO channel, the channel input−output relationship can be further described in the discrete time-domain as follows:

$$y_l(nT) = \sum_{k=1}^{N_{TX}} \sum_{m=0}^{M-1} h_{lk}(mT)x_k[(n-m)T] + n_l(nT), \quad 0 \le n \le M-1; \quad 1 \le l \le N_{RX}$$

where $x_k(n)|_{k=1,2,\ldots,N_{TX}}$ and $y_l(n)|_{l=1,2,\ldots,N_{RX}}$ represent the channel input and output time-domain signals, respectively. In the case of time-varying channels, the preceding equation can be written as $y_l(t) = \sum_{k=0}^{N_{TX}-1} h_{lk}(t,\tau)x_k(\tau) + n_l(t)$ where $h_{lk}(t,\tau)$ denotes the time-varying impulse response of the lkth channel.

The matrix form of MIMO channel output in frequency-domain sampled at single frequency ω_m can be written as $\mathbf{y}(\omega_m) = \mathbf{H}(\omega_m)\mathbf{x}(\omega_m) + \mathbf{n}(\omega_m)$. In an OFDM system, the signal

processing is inherently performed in frequency domain. The OFDM transforms a frequency-selective fading channel to a flat-fading channel when considering narrowband orthogonal subcarriers. In such system, the MIMO signal processing can be performed at each subcarrier. This is the main reason for suitability of MIMO extension to an OFDM system. When MIMO processing is performed at each subcarrier, the MIMO channel input−output relationship can be demonstrated as $\mathbf{y} = \mathbf{Hx} + \mathbf{n}$, where the channel between the transmitter and the receiver is typically modeled as a finite impulse response (FIR) filter. In this case, each tap is typically a complex-valued Gaussian random variable with exponentially decaying magnitudes. The tap delays correspond to the RMS delay spread and the channel type (e.g., low-delay spread or flat fading, high-delay spread or frequency-selective fading). There is a new realization of the channel at every transmitted packet, if the channel remains invariant for the duration of the packet; otherwise, the variation of the channel is explicitly modeled in the signal detection. As mentioned earlier, there are $N_{RX} \times N_{TX}$ paths between the transmitter and the receiver where each channel is the sum of several FIR filters with different delay spreads. The channels may or may not be correlated. The MIMO schemes can be used with non-OFDM systems when the channel is modeled as flat fading such that $y_l(nT) = \sum_{k=1}^{N_{TX}} h_{lk} x_k(nT) + n(nT)$. An important question is to what extent MIMO techniques can increase the throughput and improve the reliability of the wireless communication systems. This question can be answered by calculating the information theoretic capacity of a single-input−single-output (SISO) channel and comparing it with the capacity of single-input−multiple-output (SIMO), multiple-input−single-output (MISO), and MIMO channels.

For a memoryless SISO channel (i.e., one transmit and one receive antenna), the channel capacity is given by $C_{SISO} = \log_2(1 + \gamma|h|^2)$ where h is the normalized complex-valued gain/attenuation of a fixed wireless channel or that of a particular realization of a random channel and $\gamma = E_s/N_0$ denotes the SNR at the receive antenna port.

As the number of receive antennas increases, the statistics of channel capacity improve. Using N_{RX} receive antennas and one transmit antenna, a SIMO system is formed with a capacity given by (when the channel is unknown to the transmitter) $C_{SIMO} = \log_2(1 + \gamma \sum_{i=1}^{N_{RX}} |h_i|^2)$ where h_i is the gain of the ith channel corresponding to the ith receive-antenna. Note that increasing the value of N_{RX} results in a logarithmic increase in average channel capacity. It can be shown that knowledge of the channel at the transmitter in SIMO cases does not provide any capacity benefit.

In the case of MISO or transmit diversity, where the transmitter does not typically have knowledge of the channel, the capacity is given by $C_{MISO} = \log_2(1 + \gamma/N_{TX} \sum_{i=1}^{N_{TX}} |h_i|^2)$. It is noted from the latter equation that when the channel information is not available at the transmitter $C_{MISO} < C_{SIMO}$. The power normalization factor N_{TX} ensures that the total transmit power is uniformly distributed among the transmit antennas. Furthermore, one notes the

absence of an array gain in this case (MISO) compared to that of the receive diversity scenario where the energy of the multipath channels can be coherently combined. In addition, the MISO capacity has a logarithmic relationship with N_{TX} similar to that of a SIMO scheme. When the channel information is available at the transmitter, the capacity of a MISO system can approach to that of a SIMO system.

The use of diversity at both transmitter and receiver sides creates a MIMO system. The capacity of a MIMO system with N_{TX} transmit-antennas and N_{RX} receive antennas is expressed as $C_{MIMO} = \log_2\left(\det\left[\mathbf{I} + \gamma/N_{TX}\mathbf{H}\mathbf{H}^H\right]\right)$ where $\mathbf{I}$ is an $N_{RX} \times N_{RX}$ identity matrix and $\mathbf{H}$ is the $N_{RX} \times N_{TX}$ channel matrix. Note that both MISO and MIMO channel capacities are based on equal power and uncorrelated sources. It is demonstrated in the literature that the capacity of the MIMO channel increases linearly with $\min(N_{RX}, N_{TX})$ rather than logarithmically as in the case of MISO or SIMO channel capacity since the determinant operator yields the product of non-zero eigenvalues of its channel-dependent matrix argument, each eigenvalue characterizing the SNR over a SISO eigen-channel. It will be shown later that the overall MIMO channel capacity is the sum of capacities of each of these SISO eigen-channels. The increase in capacity is dependent on properties of the channel eigenvalues. If the channel eigenvalues decay rapidly then linear growth in capacity will not occur. However, the eigenvalues have a known limiting distribution and tend to be spaced out along the range of this distribution. Thus it is unlikely that most eigenvalues are very small, and the linear growth is indeed achieved.

The capacity of the MIMO channel can be calculated under various conditions and different assumptions. Depending on whether the receiver has perfect channel knowledge or whether the channel exhibits a flat fading or frequency-selective fading behavior, different expressions for the channel capacity can be obtained. The MIMO channel capacity can be analyzed under two different assumptions: (1) transmitter has no channel knowledge and (2) the transmitter has perfect channel knowledge through feedback from the receiver or reciprocity of the downlink and uplink channels. Let us denote by $\boldsymbol{\Xi}$ the $N_{TX} \times N_{TX}$ covariance matrix of the channel input vector $\mathbf{x}$ and let further assume that the channel is unknown to the transmitter then it can be shown that the MIMO channel capacity can be written as $C_{MIMO} = \log_2\left(\det\left[\mathbf{I} + \mathbf{H}\boldsymbol{\Xi}\mathbf{H}^H\right]\right)$ where $\mathrm{tr}(\boldsymbol{\Xi}) \leq \gamma$ ensures the total signal-power does not exceed a certain limit. It can be shown that for equal transmit power and uncorrelated sources $\boldsymbol{\Xi} = (\gamma/N_{TX})\mathbf{I}$. This is true when the channel matrix is unknown to the transmitter and the input signal is Gaussian-distributed, maximizing the mutual information. If the receiver measures and sends channel quality feedback or CSI to the transmitter, the covariance matrix $\boldsymbol{\Xi}$ is not proportional to the identity matrix, rather it is constructed from a water-filling algorithm. If one compares the capacity achieved assuming equal transmit-power and unknown channel with that of perfect channel estimation through feedback then the capacity-gain due to use of feedback is obtained.

For the independent identically distributed Rayleigh fading scenario, the linear capacity growth discussed earlier will be observed. It is shown that MIMO channel capacity can be written as $C_{MIMO} = \sum_{i=1}^{\min(N_{TX},N_{RX})} \log_2\left(1+\gamma\lambda_i^2/N_{TX}\right)$ where $\lambda_i, i = 1, 2, \ldots, \min(N_{TX}, N_{RX})$ are the non-zero eigenvalues of $\mathbf{HH}^H$. We can decompose the MIMO channel into $K \leq \min(N_{TX}, N_{RX})$ equivalent parallel SISO channels using the singular value decomposition (SVD) theorem.[14] Let $\mathbf{y} = \mathbf{Hx} + \mathbf{n}$ describe the input−output relationship of the MIMO channel where $\mathbf{y}$ is the output vector with N_{RX} components, $\mathbf{x}$ is the input vector with N_{TX} components, $\mathbf{n}$ is the additive noise vector with N_{RX} components, and $\mathbf{H}$ is the $N_{RX} \times N_{TX}$ channel matrix. Using the SVD theorem, we can show $\mathbf{H} = \mathbf{U\Sigma V}^H$. Let $\widehat{\mathbf{x}} = \mathbf{V}^H\mathbf{x}$, $\widehat{\mathbf{y}} = \mathbf{U}^H\mathbf{y}$, and $\widehat{\mathbf{n}} = \mathbf{U}^H\mathbf{n}$ denote the unitary transformation of the channel input and output and noise vectors, it can be shown that $\widehat{\mathbf{y}} = \mathbf{\Sigma}\widehat{\mathbf{x}} + \widehat{\mathbf{n}}$. Since $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices and $\mathbf{\Sigma} = \mathbf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{\min(N_{RX},N_{TX})}, 0, 0, \ldots, 0)$, it is clear that the capacity of this model is the same as the capacity of the model $\mathbf{y} = \mathbf{Hx} + \mathbf{n}$. However, $\mathbf{\Sigma}$ is a diagonal matrix with K non-zero elements on the main diagonal, thus $\widehat{y}_1 = \lambda_1\widehat{x}_1 + \widehat{n}_1, \ldots, \widehat{y}_K = \lambda_K\widehat{x}_K + \widehat{n}_K, \widehat{y}_{K+1} = \widehat{n}_{K+1}$. The latter equations are conceptually equivalent to K parallel SISO eigenchannels, each with signal power of $\lambda_i^2, i = 1, 2, \ldots, \min(N_{TX}, N_{RX})$. As a result, the MIMO channel capacity can be rewritten in terms of the eigenvalues of the input signal covariance matrix $\mathbf{\Xi}$.

When the channel knowledge is available at the transmitter and receiver then $\mathbf{H}$ is known, and we can optimize the capacity over $\mathbf{\Xi}$ subject to the power constraint $\mathrm{tr}(\mathbf{\Xi}) \leq \gamma$. It is shown in the literature that the optimal $\mathbf{\Xi}$ in this case exists and is known as water-filling solution. The channel capacity in this case is given by

$$C = \sum_{k=1}^{K} \log_2\left(\eta\lambda_i^2\right)^+, \quad \gamma = \sum_{k=1}^{K}\left(\eta - \lambda_i^{-2}\right)^+$$

where $(x)^+ = x \forall x \geq 0, (x)^+ = 0 \forall x < 0$ and η is a nonlinear function of eigenvalues of the channel input covariance matrix. The effect of various channel conditions on the channel capacity has been extensively studied in the literature. For example, increasing the LoS signal strength at fixed SNR reduces capacity in Rician channels [22,24]. This can be explained in terms of the channel matrix rank or through various eigenvalue properties. The issue of correlated fading is of considerable importance for implementations where the

[14] The concept of decomposition of an $N \times N$ Hermitian matrix in terms of quadratic product of $N \times N$ unitary matrix composed of eigenvectors and $N \times N$ diagonal matrix of eigenvalues can be generalized to $M \times N$ complex-valued matrices of rank K. If $\mathbf{A}$ is a $M \times N$where$(M > N)$ complex-valued matrix of rank K then $\mathbf{A} = \mathbf{U\Sigma V}^H$ denotes the singular value decomposition of $\mathbf{A}$. The $M \times M$ unitary matrix $\mathbf{U}$ is composed of the eigenvectors of $\mathbf{AA}^H$, that is, $\mathbf{U} = (\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_m})$, where $\mathbf{AA}^H\mathbf{u}_i = \sigma_i^2\mathbf{u}_i$. The $N \times N$ unitary matrix $\mathbf{V}$ is composed of eigenvectors of $\mathbf{A}^H\mathbf{A}$, that is, $\mathbf{V} = (\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_n})$, where $\mathbf{A}^H\mathbf{v}_i = \sigma_i^2\mathbf{v}_i$ The elements of the $M \times N$ are the square roots of the eigenvalues of matrix $\mathbf{A}^H\mathbf{A}$o as the singular values of matrix $\mathbf{A}$ may be written as $\mathbf{A} = \sum_{i=1}^{K} \sigma_i\mathbf{u}_i\mathbf{v}_i^H$singular values of matrix $\mathbf{A}$ is the rank of $\mathbf{A}$. The singular values of matrix $\mathbf{A}$ are positive real numbers which satisfy $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_K > 0$ [22].

antennas are required to be closely spaced. The optimal water-filling allocation strategy is obtained when the power allocated to each spatial subchannel is non-negative.

In the design of wireless communication systems, the main objective is to exploit the transmission schemes whose performance can approach the channel capacity as much as possible. Therefore, it is important to understand the underlying concepts and various information theoretic definitions of channel capacity and what can be pragmatically achieved under realistic channel conditions and transceiver implementations. Let us begin our concise study with the most generic definition of channel capacity. We denote the input and output of a memoryless SISO wireless channel with the random variables X and Y, respectively, the channel capacity is defined as $C = \max_{p(x)} I(X;Y)$ where $I(X;Y)$ represents the mutual information between X and Y. Shannon's theorem [19] provides an operational meaning to the definition of the instantaneous capacity as the number of bits that can be transmitted reliably over the channel with vanishing probability of error. The mutual information is maximized with respect to all possible transmit signal statistical distributions $p(x)$. Mutual information is a measure of the amount of information that one random variable contains about another variable. The mutual information between X and Y can also be written as $I(X;Y) = H(Y) - H(Y|X)$ where $H(Y|X)$ represents the conditional entropy between the random variables X and Y. The entropy of a random variable can be described as the measure of uncertainty in the random variable or the amount of information required on the average to describe the random variable. Thus the mutual information representation of channel capacity can be described as the reduction in the uncertainty of one random variable due to the knowledge of the other. Note that the mutual information between X and Y depends on the properties of the channel through a channel matrix $\mathbf{H}$ and the properties of X through the probability distribution of X [22].

Throughout this section, it is assumed that the channel matrix $\mathbf{H}$ is random and that the receiver has perfect channel knowledge. It is also assumed that the channel is memoryless, that is, for each use of the channel an independent realization of $\mathbf{H}$ is drawn. This means that the capacity can be computed as the maximum of the mutual information as defined earlier. The results are also valid when $\mathbf{H}$ is generated by an ergodic process because as long as the receiver observes the $\mathbf{H}$ process, only the first order statistics are needed to determine the channel capacity.

The ergodic (mean) capacity of a random channel with $N_{RX} = N_{TX} = 1$ and an average transmit power constraint P_T can be expressed as $C_{ergodic} = E_{\mathbf{H}}\left\{\max_{p(x):P \le P_T} I(X;Y)\right\}$ where P is the average power of a single codeword, transmitted over the channel, and $E_{\mathbf{H}}\{.\}$ denotes the expectation over all channel realizations. Compared to the generic definition, the capacity of the channel is now defined as the maximum of the mutual information between the input and the output over all statistical distributions on the input that satisfy the power constraint. In general,

the capacity of a random MIMO channel with power constraint P_T can be expressed as $C_{ergodic} = E_{\mathbf{H}}\left\{\max_{p(x):\mathrm{tr}(\Phi)\le P_T} I(X;Y)\right\}$ where $\mathbf{\Phi} = E\left[\mathbf{x}\mathbf{x}^H\right]$ is the covariance matrix of the transmit signal vector $\mathbf{x}$. The total transmit power is limited to P_T irrespective of the number of transmit antennas. For a fading channel the channel matrix $\mathbf{H}$ is a stochastic process; thus the associated channel capacity $C(\mathbf{H})$ is a random variable. In this case, the ergodic channel capacity is defined as the average of instantaneous channel capacity over the distribution of $\mathbf{H}$. The ergodic channel capacity of the MIMO transmission scheme is given by

$$C_{ergodic} = E\left\{\max_{\mathrm{tr}(\mathbf{\Phi})=N_{TX}} \log_2\left(\det\left[\mathbf{I} + \gamma/N_{TX}\mathbf{H}\mathbf{\Phi}\mathbf{H}^H\right]\right)\right\}$$

where $\mathbf{\Phi}$ denotes the $N_{TX} \times N_{TX}$ covariance matrix of the channel input vector $\mathbf{x}$. According to information theoretic concepts, this capacity cannot be achieved unless channel coding is employed across an infinite number of independently fading blocks. Let us focus on the case of perfect CSI at the receiver side and no CSI at the transmitter side, which implies that the maximization of latter equation is now more restricted than in the previous case. Nevertheless, it has been shown in the literature that the optimal signal covariance matrix must be chosen according to $\mathbf{\Phi} = \mathbf{I}$. This means that the antennas should transmit uncorrelated streams with the same average power. With this result, the ergodic MIMO channel capacity reduces to [15]

$$C_{ergodic} = E\left\{\log_2\left(\det\left[\mathbf{I} + \gamma/N_{TX}\mathbf{H}\mathbf{H}^H\right]\right)\right\}$$

It is obvious that this is not the Shannon capacity in a true sense, since as mentioned earlier one with perfect channel knowledge at the transmitter can choose a signal covariance matrix that outperforms $\mathbf{\Phi} = \mathbf{I}$ case. Nevertheless, we refer to the preceding expression as the ergodic channel capacity with CSI at the receiver and no CSI at the transmitter.

The capacity under channel ergodicity is defined as the average of the maximum value of the mutual information between the transmitted and the received signals, where the maximization is carried out with respect to all possible transmit signal statistical distributions. Another measure of channel capacity that is frequently used is outage capacity. With outage capacity, the channel capacity is associated to an outage probability. Capacity is treated as a random variable which depends on the channel instantaneous response and remains constant during the transmission of a finite-length coded block of information. If the channel capacity falls below the outage capacity, there is no possibility that the transmitted block of information can be decoded with no errors, no matter which coding scheme is employed. The probability that the capacity is less than the outage capacity denoted by C_{outage} is ρ. This can be expressed in mathematical terms by $p(C < C_{outage}) = \rho$. In this case the latter expression represents an upper bound since there is a finite probability ρ that the channel capacity is less than the outage capacity. It can also be written as a lower bound, representing the case where there is a finite probability $(1 - \rho)$ that the channel capacity is higher than

C_{outage}, which means $p(C > C_{outage}) = 1 - \rho$. In other words, since the MIMO instantaneous channel capacity is a random variable, it is meaningful to consider its statistical distribution, thus a useful measure of its statistical behavior is the outage capacity. Outage analysis quantifies the level of performance (in this case capacity) that is guaranteed with a certain level of reliability. The $\rho\%$ outage capacity $C_{outage}(\rho)$ is defined as the information rate that is guaranteed for $(100 - \rho)\%$ of the channel realizations ($p(C < C_{outage}) = \rho$). The outage capacity is often a more relevant measure than the ergodic channel capacity, because it describes in some way the quality of the channel. This is due to the fact that the outage capacity measures the probabilistic distribution of the instantaneous rate supported by the channel. Thus, if the rate supported by the channel is spread over a wide range, the outage capacity for a fixed probability level may become small, whereas the ergodic channel capacity may be high [15].

4.1.9.2 Single-User and Multi-user MIMO

Single-user MIMO (SU-MIMO) techniques are point-to-point schemes that improve channel capacity and reliability through the use of space-time/space-frequency codes (transmit/receive diversity) in conjunction with spatial multiplexing schemes. In an SU-MIMO transmission, the advantage of MIMO processing is obtained from the coordination of processing among all the transmitters or receivers. In the multi-user channel, on the other hand, it is usually assumed that there is no coordination among the users. As a result of the lack of coordination among users, uplink and downlink multi-user MIMO channels are different. In the uplink scenario, users transmit to the base station over the same channel. The challenge for the base station is to separate the signals transmitted by the users, using array processing or multi-user detection methods. Since the users are not able to coordinate with each other, there is not much that can be done to optimize the transmitted signals with respect to each other. If some channel feedback is allowed from the transmitter back to the users, some coordination may be possible, but it may require that each user know all the other users' channels rather than only its own. Otherwise, the challenge in the uplink is mainly in the processing done by the base station to separate the users. In the downlink channel, where the base station simultaneously transmits to a group of users over the same channel, there is some inter-user interference for each user which is generated by the signals transmitted to other users. Using multi-user detection techniques, it may be possible for a given user to overcome the multiple access interference, but such techniques are often extremely complicated for use at the receivers. Ideally, one would like to mitigate the interference at the transmitter by carefully designing the transmit signal. If CSI is available at the transmitter, it is aware of what interference is caused for each user by the signal it is transmitted to other users. The inter-user interference can be mitigated by beamforming or the use of dirty paper codes. In general, single-user and multi-user MIMO schemes are compared as follows [15]:

- SU-MIMO is a point-to-point link with predictable link capacity, whereas MU-MIMO channel is a broadcast channel (BC) in the downlink direction and a multiple access channel (MAC) in the uplink direction whose link-level data rates are characterized in terms of capacity regions.
- Multi-layer SU-MIMO schemes offer layer/stream diversity in the sense that if one stream has a poor SNR, the system will not necessarily experience an outage, whereas in the same situation, a MU-MIMO system will be in outage. This is because in MU-MIMO schemes, users have typically an equal target data rate and symbol error rate on their respective links, while in SU-MIMO systems, only the sum rate of the overall link is considered since all streams are delivered to same user.
- The MU-MIMO schemes suffer from near-far problem due to significant difference between the path losses experienced by each user, resulting in large deviation in the SINR of the corresponding user links. This would benefit the users with better channel conditions, while there is no near-far problem in SU-MIMO systems. The near-far problem in MU-MIMO systems may be alleviated via appropriate grouping of the users with similar channel conditions.
- The use of cooperative collocated transmit antennas in SU-MIMO schemes can facilitate the encoding at the transmitter and decoding at the receiver. In contrast, the users in a MU-MIMO scheme can cooperate in encoding at the base station in the downlink and decoding in the uplink; however, the users cannot cooperate in decoding in the downlink or encoding in the uplink directions.
- The capacity of the downlink and uplink is theoretically identical in the SU-MIMO systems (given the same transmit power and the perfect channel knowledge in the transmitter and the receiver); however, the capacities of the MU-MIMO BC and MAC are not identical.
- The capacity of the SU-MIMO schemes is less impacted by lack of CSI at the transmitter, whereas the capacity of the MU-MIMO BC significantly suffers from lack of CSI at the transmitter.
- SU-MIMO suffers from limited exploitation of multi-user diversity. The number of spatial dimensions is limited by number of antennas at the UE. There is a potential that spatial dimensions are wasted, if the UEs have a smaller number of antennas compared to the base station.
- MU-MIMO more efficiently exploits the multi-user diversity since all spatial dimensions which are supported by the base station can be exploited. It will achieve capacity gain, if UEs have a smaller number of antennas relative to the base station. Stronger spatial dimensions are exploited, particularly in the case of low-rank channel. The utilized spatial dimensions may be weak in the case of low-rank channel due to spatial correlation.

The advantages/disadvantages of SU-MIMO and MU-MIMO schemes are summarized in Table 4.21.

Table 4.21: Comparison of SU-MIMO and MU-MIMO schemes.

	SU-MIMO	MU-MIMO
Advantages	High user throughput High peak data rates	High system capacity Full exploitation of multiuser diversity
Disadvantages	Multiple transmit antennas at the base station are not fully exploited Multiuser diversity is not fully exploited	Degradation of peak data rates due to interuser interference

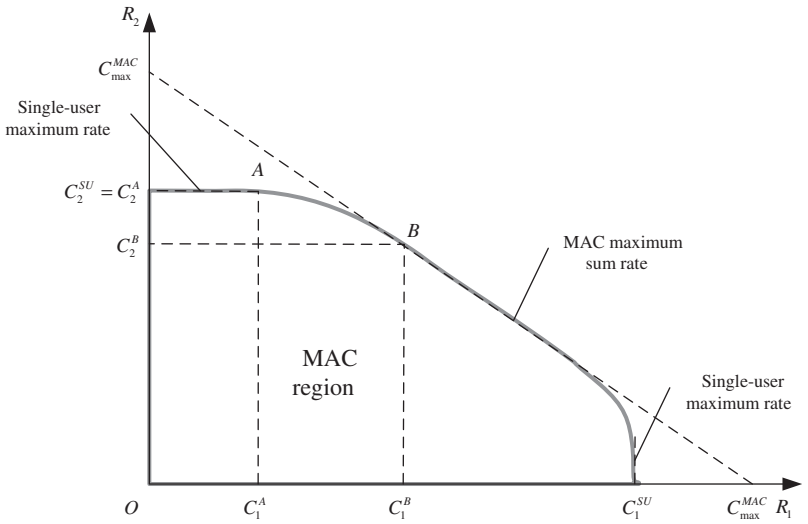


Figure 4.60
Capacity region of MU-MIMO BC with two users compared to SU-MIMO.

An important metric for measuring the performance of any communication channel is the information theoretic capacity. In an SU-MIMO channel, the capacity is the maximum amount of information that can be transmitted as a function of available bandwidth given a constraint on transmitted power. In SU-MIMO channels, it is common to assume that the total power distributed among all transmit antennas is limited. For the multi-user MIMO channel, the problem is somewhat more complex. Given a constraint on the total transmit power, it is possible to allocate varying fractions of that power to different users in the network; thus for any value of total power, different information rates are obtained. The result is a capacity region shown in Fig. 4.60 for two-user MU-MIMO channel. The maximum capacity for user 1 is achieved when 100% of the power is allocated to user 1, and for user 2, the maximum capacity is also obtained when it is allocated the full power. For every possible power distribution, there is an achievable information rate, which results in the

capacity regions depicted in the figure. Two regions are shown in Fig. 4.60, the larger one for the case where both users have roughly the same maximum capacity (similar channel conditions), and the other region for a case where one of the users has much better channel condition than the other user. For N_{user} users, the capacity region is characterized by an N_{user}-dimensional hyper-region.

Let us use a simple MU-MIMO system model to demonstrate how the sum rate of the system is calculated. As shown in Fig. 4.61, the transmit vector $\mathbf{x}$ can be expressed as the weighted sum (precoded) of the input data symbols $s_k|_{k=1,2,\ldots,N_{user}}$ as follows $\mathbf{x} = \sum_k \mathbf{p}_k s_k = \mathbf{Ps}$ where $N_{user} \times 1$ vector $\mathbf{s} = \left(s_1, s_2, \ldots, s_{N_{user}}\right)^T$ denotes the data symbols from N_{user} users and $\mathbf{P} = \left(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{N_{user}}\right)$ is the precoding matrix comprising N_{user} precoding vectors. It is assumed that the finite transmit power at the transmitter can be calculated as follows $P_{TX} = E\left\{\mathbf{x}^H\mathbf{x}\right\}$. The kth complex-valued output of the system can be written as $\mathbf{y}_k = \mathbf{H}_k\mathbf{x}_k + \mathbf{n}_k \in \mathbb{C}^N$ where N denotes the dimension of vector $\mathbf{y}$. The kth branch user data can be detected using a linear minimum mean squared error (MMSE) receiver as follows $\hat{s}_k = \mathbf{w}_k^T\mathbf{x}_k \in \mathbb{C}$ in which the MMSE weighting matrix is given by

$$\mathbf{w}_k = \mathbf{H}_k^*\mathbf{P}^*\mathbf{P}^T\mathbf{H}_k^T + \frac{N_{user}}{P_{TX}}\mathbf{I}_N{}^{-1}\mathbf{H}_k^*\mathbf{p}_k^*$$

It can be further shown that the SINR at the kth output is given by

$$\gamma_k = \frac{\left|\mathbf{w}_k^T\mathbf{H}_k\mathbf{p}_k\right|^2}{\left\|\mathbf{w}_k\right\|_2^2 N_{user}/P_{TX} + \sum_{i(i\neq k)}\left|\mathbf{w}_k^T\mathbf{H}_k\mathbf{p}_i\right|^2}$$

The sum rate of the system is given as $R_{sum} = \sum_k \log_2(1+\gamma_k)$.

In the uplink of a multi-user MIMO system, the received signal at the gNB can be written as $\mathbf{y} = \sum_{k=1}^{N_{user}} \mathbf{H}_k^H\mathbf{x}_k + \mathbf{n}$ where $\mathbf{x}_k$ is the $N_{TX_k} \times 1$ transmitted signal vector of the kth UE with N_{TX_k} transmit antennas, $\mathbf{H}_k \in \mathbb{C}^{N_{TX_k} \times N_{RX}}$ denotes the flat-fading channel matrix from the kth user to the gNB and $\mathbf{n} = (n_1, n_2, \ldots, n_{N_{RX}}), n_k \sim N(0,1)$ is an independent and identically distributed additive white Gaussian noise vector at the gNB. We assume that the receiver k has perfect and instantaneous knowledge of the channel matrix $\mathbf{H}_k$. Note that the gNB is equipped with N_{TX} transmit and N_{RX} receive antennas.

In the downlink, the received signal at the kth receiver can be written as $\mathbf{y}_k = \mathbf{H}_k\mathbf{x} + \mathbf{n}_k \forall k = 1, 2, \ldots, N_{user}$ where $\mathbf{H}_k \in \mathbb{C}^{N_{RX_k} \times N_{TX}}$ is the downlink channel, and $\mathbf{n}_k \in \mathbb{C}^{N_{RX_k} \times N_{TX}}$ is the complex-valued additive Gaussian noise at the kth receiver. We assume that each receiver also has perfect and instantaneous knowledge of its own channel matrix $\mathbf{H}_k$. The transmitted signal $\mathbf{x}$ is a function of the multiple users' information data, that is, $\mathbf{x} = \sum_{k=1}^{N_{user}} \mathbf{x}_k$ where $\mathbf{x}_k$ is the signal carrying kth user's message with covariance matrix $\mathbf{\Omega}_k = E\{\mathbf{x}_k\mathbf{x}_k^H\}$. The power allocated to the kth user is given by $\rho_k = \mathrm{tr}\{\mathbf{\Omega}_k\}$. Under a

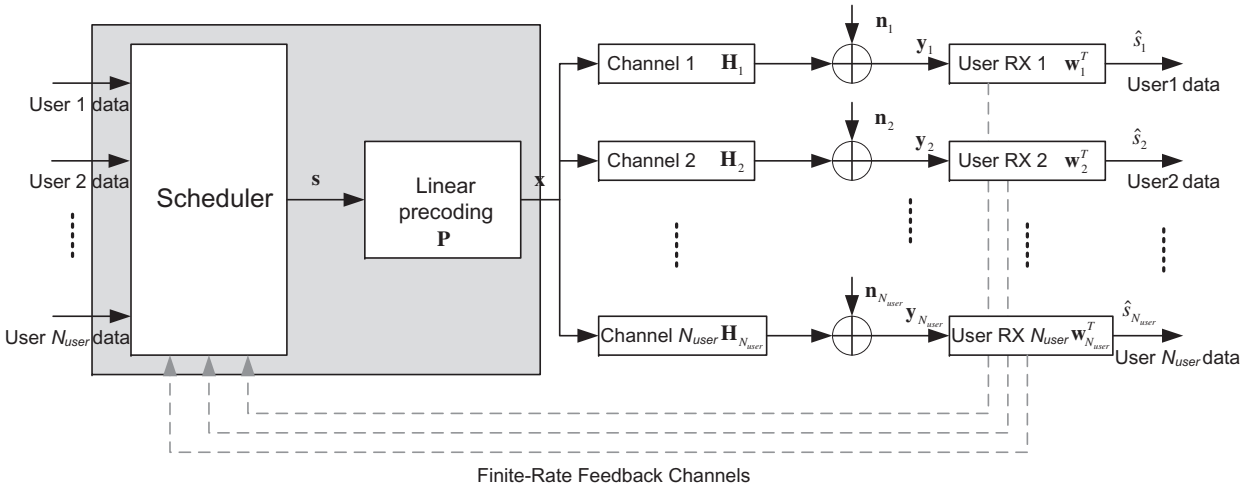


Figure 4.61
MU-MIMO BC model [15].

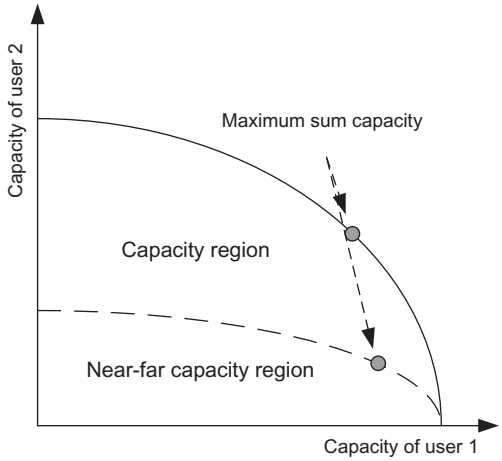


Figure 4.62
An example illustration of capacity region.

sum power constraint at the gNB, the power allocation needs to maintain $\sum_{k=1}^{N_{user}} \rho_k \leq P_{total}$. Assuming a unit variance for the noise, it can be shown that the capacity region for a given matrix channel realization can be written as [15]

$$C_{DL} = \bigcup_{\left(\rho_1,\rho_2,\ldots,\rho_{N_{user}} \mid \sum \rho_k \leq P_{total}\right)} \left\{ \left(R_1,R_2,\ldots,R_{N_{user}}\right) \in \mathbb{R}^{+N_{user}}, R_l \leq \log_2 \frac{\det\left[\mathbf{I}+\mathbf{H}_i\left(\sum_{j\geq i}\boldsymbol{\Omega}_j\right)\mathbf{H}_i^H\right]}{\det\left[\mathbf{I}+\mathbf{H}_i\left(\sum_{j>i}\boldsymbol{\Omega}_j\right)\mathbf{H}_i^H\right]} \right\}$$

where $\mathbb{R}^{+N_{user}}$ is the N_{user}-dimensional set of positive real numbers. The preceding equation may be optimized over each possible user ordering. Although difficult to realize in practice, the computation of the capacity region can be simplified using the assumption that the downlink capacity region can be calculated through the union of regions of the dual MAC with all uplink power allocation vectors meeting the sum power constraint. The fundamental effect of the use of multiple antennas at either the gNB or the user terminals in increasing the channel capacity is best realized by examining how the sum capacity, that is, the point obtained by the maximum $\sum_{k=1}^{N_{user}} R_k$ in the capacity region, scales with the number of active users (see Fig. 4.62).

An efficient UE pairing scheme is required at the gNB to choose the correct pair of UEs for transmission in MU-MIMO systems. This pairing scheme is required to maintain minimal interference among scheduled UEs in MU-MIMO transmission. A proper pairing scheme can be designed by maximizing the chordal distance[15] between the feedback precoding

[15] The asymptotic performance of a coding scheme is dominated by the shortest distance between any pair of codewords. The relevant distance measure between two codewords $\mathbf{x}_1$ and $\mathbf{x}_2$ of an orthogonal code for a non-coherent MIMO system is the chordal distance defined as $d^2(\mathbf{x}_1, \mathbf{x}_2) = M - \left\|\mathbf{x}_1\mathbf{x}_2^H\right\|_F^2$.

matrices of the UEs. The chordal distance between two matrices is given in [15] and represented by $d_c(\mathbf{p}_i, \mathbf{p}_j) = \frac{1}{\sqrt{2}} \left\| \mathbf{p}_i \mathbf{p}_i^H - \mathbf{p}_j \mathbf{p}_j^H \right\|_F$ where $\|.\|_F$ denotes the Frobenius norm of the matrix. The chordal distance generalizes the distance between two points on the unit sphere through an isometric embedding from complex Grassmann manifold $Gr(N_{TX}, N_l)$ to the unit sphere. Assuming an infinite number of UEs served by the current gNB, the kth UE with reported precoding matrix $\mathbf{p}_k$ will be paired with the mth UE, where the mth UE reports precoding matrix $\mathbf{p}_m$ and the chordal distance between precoding matrices is maximized. With the maximized chordal distance criterion, $\mathbf{p}_m$ stays in the anti-polar position of $\mathbf{p}_k$; hence, $\left\| \mathbf{H}_k \mathbf{p}_m \right\|^2$ is minimized, yielding the minimized inter-user interference. Therefore, the UE pairing scheme for MU-MIMO transmission in practical systems is designed to find the best match between two UEs (e.g., the mth UE and the kth UE) based on the reported precoding matrices and the following criterion $\mathbf{p}_m = \arg \max_{\forall \mathbf{p}_i \in \mathcal{P}_{UE}} d_c(\mathbf{p}_i, \mathbf{p}_k)$ with $\mathcal{P}_{UE}$ representing the pool containing all reported precoding matrices at a certain gNB.

4.1.9.3 Analog, Digital, and Hybrid Beamforming

Large antenna arrays and beamforming play an important role in 5G implementations since both base stations and devices can accommodate a larger number of antenna elements at mmWave frequencies. Aside from a higher directional gain, these antenna types offer complex beamforming capabilities. This allows increasing the capacity of cellular networks by improving the signal-to-interference ratio through direct targeting of user groups. The narrow transmit-beams simultaneously lower the amount of interference in the radio propagation environment and make it possible to maintain sufficient signal power at the receiver terminal at larger distances in rural areas. An important prerequisite for any beamforming architecture is a phase coherent signal, which means that there is a defined and stable phase relationship between all RF carriers. A fixed phase offset between the carriers can be used to steer the main antenna lobe to a desired direction. A major difference between NR and LTE is the support for beamformed control channels, which resulted in different reference signal design for each control channel. The NR physical channels and signals, including those used for control and synchronization, have all been designed to support beamforming. The CSI for operation with a large number of antennas can be obtained via CSI reports from the devices based on transmission of CSI-RSs in the downlink, as well as using uplink measurements exploiting channel reciprocity.

4.1.9.3.1 Analog Beamforming

Analog beamforming typically relies on conditioning the amplitude and phase of the signals that feed the antenna array. The combination of these two factors is used to improve sidelobe suppression or steering nulls. Phase and amplitude for each antenna element are combined by applying a complex-valued weighting factor to the signal that is fed to the

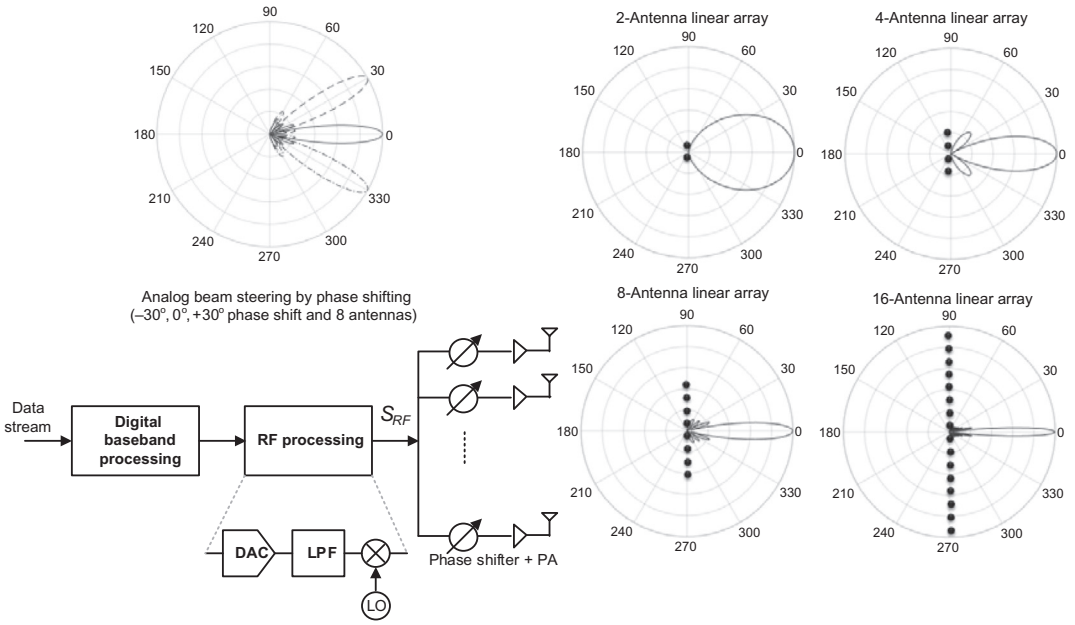


Figure 4.63
Analog beamforming transmitter architecture.

corresponding antenna. Fig. 4.63 shows a basic implementation of an analog beamforming transmitter architecture. This architecture consists of only one RF chain and multiple phase shifters that feed an antenna array. The phased arrays have been used in practical systems (e.g., radar systems) for the past several decades. Beam steering was often carried out with a selective RF switch and fixed phase shifters. This concept is still used in modern communication systems using advanced hardware and improved precoding techniques. These enhancements enable separate control of the phase of each element. Unlike traditional passive architectures, the beam can be steered not only to discrete but virtually any angle using active beamforming antennas. Analog beamforming is performed in the analog domain at RF frequencies or an intermediate frequency. However, implementing multi-stream transmission with analog beamforming is a highly complex task. In order to calculate the phase shifts, a uniformly spaced linear array with element spacing is assumed. Considering the receive scenario, the antenna array must be in the far field of the incoming signal, so that the arriving wave front is approximately planar. If the signal arrives at an angle θ relative to the antenna boresight, the wave must travel an additional distance $d \sin \theta$ to arrive at each successive element. This translates to an element specific delay which can be converted to a frequency-dependent phase shift of the signal as $\Delta\varphi = 2\pi d(\sin \theta)/\lambda$. The frequency dependency translates into an effect called beam unevenness. The main lobe of an antenna array at a defined frequency can be steered to a certain angle using phase offsets calculated by the latter equation (see Fig. 4.63). If the

antenna elements are now fed with a signal of a different frequency, the main lobe will swerve by a certain angle. Since the phase relations were calculated with a certain carrier frequency in mind, the actual angle of the main lobe shifts according to the current frequency. Radar applications with large bandwidths in particular suffer inaccuracies due to this effect. The latter equation can be expressed in time domain using time delays instead of frequency offsets $\Delta\tau = d/c \sin\theta$. This means that the frequency dependency is eliminated, if the setup is fitted with delay lines instead of phase shifters. The performance of the analog architecture can be further improved by additionally changing the magnitude of the signals feeding the antennas [56].

Analog signal processing typically implies that beamforming is carried out on a per-carrier basis. For the downlink transmission, this implies that it is not possible to frequency-multiplex beamformed transmissions to devices located in different directions relative to the base station. In other words, beamformed transmissions to different devices located in different directions must be separated in time.

4.1.9.3.2 Digital Beamforming

While analog beamforming is generally restricted to one RF chain even when using large number of antenna arrays, digital beamforming in theory supports as many RF chains as there are antenna elements. If suitable precoding is performed in the digital-domain baseband, this would yield higher flexibility regarding the transmission and reception. The additional degree of freedom can be leveraged to perform advanced techniques such as multi-beam MIMO. These advantages result in the highest theoretical performance possible compared to other beamforming architectures. Fig. 4.64 illustrates the high-level digital beamforming transmitter architecture with multiple RF chains. Uneven beam is a problem

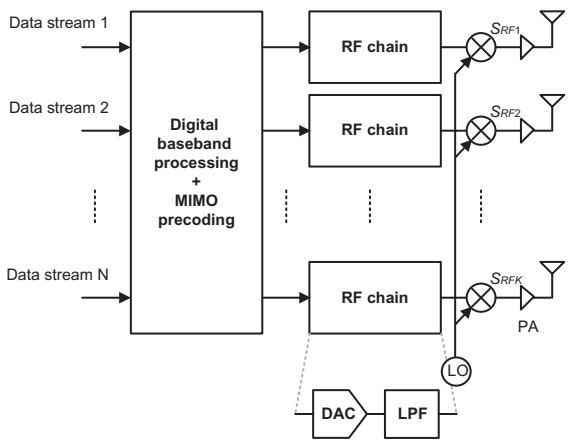


Figure 4.64
Digital beamforming transmitter architecture.

of analog beamforming architectures using phase shifters. This is a drawback considering 5G plans to make use of large bandwidths in the mmWave bands. Digital control of the RF chain enables optimization of the phases according to the frequency over a large band. Nonetheless, digital beamforming may not always be ideally suited for practical implementations of 5G applications. The very high complexity and requirements regarding the hardware may significantly increase cost, energy consumption, and complicate integration in mobile devices. Digital beamforming is better suited for use in base stations, since performance outweighs mobility in this case. Digital beamforming can accommodate multi-stream transmission and serve multiple users simultaneously, which is a key driver of the technology [56].

Multiple antennas at the transmitter and receiver can be used to achieve array and diversity gain instead of capacity gain. In this case, the same symbol weighted by a complex-valued scale factor is sent from each transmit antenna, so that the input covariance matrix has unit rank. This scheme is referred to as beamforming. It must be noted that there are two conceptually and practically different classes of beamforming: (1) direction-of-arrival beamforming (i.e., adjustment of transmit or receive antenna directivity) and (2) eigen-beamforming (i.e., a mathematical approach to maximize signal power at the receive antenna based on certain criterion). In this section, we only consider eigen-beamforming schemes.

A classic eigen-beamforming scheme usually performs linear, single-layer, complex-valued weighting on the transmit symbols such that the same signal is transmitted from each transmit antenna using appropriate weighting factors. In this scheme, the objective is to maximize the signal power at the receiver output. When the receiver has multiple antennas, the single-layer beamforming cannot simultaneously maximize the signal power at the receive antennas; hence, precoding is used for multi-layer beamforming in order to maximize the throughput of a multi-antenna system. Precoding is a generalized beamforming scheme to support multi-layer transmission in a MIMO system. Using precoding, multiple streams are transmitted from the transmit antennas with independent and appropriate weighting per each antenna such that the throughput is maximized at the receiver output.

Let us begin our concise study of eigen-beamforming and MIMO precoding using a simplified model, where we have two transmit antennas at the base station and a UE with a single receive antenna. The goal is to find the complex-valued precoding weights such that the SNR at the receiver is maximized. The channel in this example is a vector $\mathbf{h} = (h_1, h_2)$ where h_1 and h_2 are channel coefficients. It can be shown that the complex-valued weighting factors p_1 and p_2 can be calculated as shown in Fig. 4.65. This example illustrates the concept of digital precoding and how in theory the weighting vectors/matrices are calculated to maximize the SNR at the receiver.

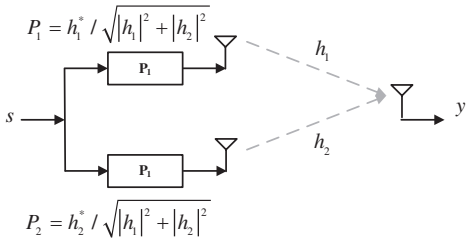


Figure 4.65
Concept of the digital precoding.

In an SU-MIMO system, the identity matrix precoding (for open-loop) and SVD precoding (for closed-loop) can be used to achieve link-level MIMO channel capacity. In addition, random unitary precoding can achieve the open-loop MIMO channel capacity with no signaling overhead in the uplink. The SVD precoding, on the other hand, has been shown to achieve the MIMO channel capacity when CSI is known at the transmitter. In a precoded SU-MIMO system with N_{TX} transmit antennas and N_{RX} receive antennas, the input−output relationship can be described as $\mathbf{y} = \mathbf{HWs} + \mathbf{n}$ where $\mathbf{s} = (s_1, s_2 \ldots, s_M)^T$ is an $M \times 1$ vector of normalized complex-valued modulated symbols, $\mathbf{y} = \left(y_1, y_2, \ldots, y_{N_{RX}}\right)^T$ and $\mathbf{n} = \left(n_1, n_2, \ldots, n_{N_{RX}}\right)^T$ are the $N_{RX} \times 1$ vectors of received signal and noise, respectively, $\mathbf{H}$ is the $N_{RX} \times N_{TX}$ complex-valued channel matrix, and $\mathbf{W}$ is the $N_{TX} \times M$ linear *precoding* matrix.

In the receiver, a hard-decoded symbol vector $\hat{\mathbf{s}}$ is obtained by decoding the received vector $\mathbf{y}$ using a vector decoder, assuming perfect knowledge of the channel and the optimal selection of precoding matrices. We assume that the entries of $\mathbf{H}$ are independent and identically distributed according to $\mathbb{C}(0, 1)$ (complex-valued normal distribution) and the entries of noise vector $\mathbf{n}$ are independent and identically distributed according to $\mathbb{C}(0, N_0)$. The input vector $\mathbf{s}$ is assumed to be normalized, thus $E[\mathbf{ss}^H] = \mathbf{I}$ where $\mathbf{I}$ is an identity matrix. Let us further assume that precoding matrix $\mathbf{W}$ is unitary, thus $\mathbf{WW}^H = \mathbf{I}$. The receiver selects a precoding matrix $\mathbf{W}_i, i = 1, 2, \ldots, N_{codebook}$ from a finite set of quantized precoding matrices $\boldsymbol{\Omega} = \{\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_{N_{codebook}}\}$ and sends the index of the chosen precoding matrix back to the transmitter over a low-delay feedback channel. There are two important issues concerning the above precoding scheme: (1) optimal selection criterion for choosing a precoding matrix from set $\boldsymbol{\Omega}$ and (2) design of codebook $\boldsymbol{\Omega}$. The matrix $\mathbf{W}_i, i = 1, 2, \ldots, N_{codebook}$ can be selected from $\boldsymbol{\Omega}$ by using either of the following optimization criteria [75]: (1) minimizing the trace of the mean squared error (MMSE-trace selection), (2) minimizing the determinant of the mean squared error (MMSE-determinant selection), (3) maximizing the minimum singular value of $\mathbf{HW}$ (singular value selection), (4) maximizing the instantaneous capacity (capacity selection), or (5) maximizing the minimum received symbol vector distance (minimum distance selection). The above selection criteria may be evaluated at the receiver using

a full search over all matrices in $\mathbf{\Omega}$. Using distortion functions based on the selection criteria, it can be shown that the codebook $\mathbf{\Omega}$ is designed using Grassmannian subspace packing [75]. If MMSE-trace, singular value, or minimum distance selection is used, the codebook is designed such that $\varepsilon = \min_{\mathbf{W_i} \neq \mathbf{W_j}} \left\| \mathbf{W}_i \mathbf{W}_i^H - \mathbf{W}_j \mathbf{W}_j^H \right\|_2$[16] is maximized. If MMSE-determinant or capacity selection optimization method is used, the codebook is designed such that $\varepsilon = \min_{\mathbf{W_i} \neq \mathbf{W_j}} \arccos \left| \det\left(\mathbf{W}_i^H \mathbf{W}_j \right) \right|$ is maximized.

The precoding matrices (MIMO codebooks) are designed based on a trade-off between performance and complexity. The following are some desirable properties of the codebooks:

1. Low-complexity codebooks can be designed by choosing the elements of each constituent matrix or vector from a small binary set, for example, a four alphabet (± 1, $\pm j$) set, which eliminates the need for matrix or vector multiplication. In addition, nested property of the codebooks can further reduce the complexity of CQI calculation when rank adaptation is performed.
2. Base station may perform rank overriding which results in significant CQI mismatch, if the codebook structure cannot adapt to it. A nested property with respect to rank overriding can be exploited to mitigate the mismatch effects.
3. Power amplifier balance is taken into consideration when designing codebooks with constant modulus property, which may eliminate unnecessary increase in PAPR.
4. Good performance for a wide range of propagation scenarios, for example, uncorrelated, correlated, and dual-polarized channels, is expected from the codebook design algorithms. A DFT-based codebook is optimal for linear array with small antenna spacing since the vectors match with the structure of the transmit array response. In addition, with an optimal selection of the matrices and the entries of the codebook (rotated block diagonal structure), significant gains can be obtained in dual-polarized scenarios.
5. Low feedback and signaling overhead are desirable from operation and performance perspective.
6. Low memory requirement is another design consideration for the MIMO codebooks.

Let us consider multi-user MIMO systems and briefly study how precoding is applied in those scenarios. In the downlink direction of a precoded MU-MIMO system (alternatively known as BC in the literature) with N_{TX} transmit antennas at the base station and one receive antenna at the kth mobile station the input−output relationship can be written as $y_k = \mathbf{h}_k^H \mathbf{x} + n_k, k = 1, 2, \ldots, N_{user}$, where $\mathbf{x} = \sum_{i=1}^{N_{user}} s_i \mathbf{w}_i$ is the $N_{TX} \times 1$ vector of weighted transmitted symbols s_i, y_k and n_k are the received signal and noise, respectively, $\mathbf{h}_k$ is the kth $N_{TX} \times 1$ channel vector, where matrix $\mathbf{H} = \left(\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{N_{user}} \right)^T$ is the $N_{user} \times N_{TX}$

[16] The Euclidean norm of square matrix $\mathbf{A}$ is defined as $\|\mathbf{A}\|_2 = \sup_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|_2$. The spectral norm of matrix $\mathbf{A}$ is the largest singular value of $\mathbf{A}$ or the square root of the largest eigenvalue of the positive-semi-definite matrix $\mathbf{A}^H\mathbf{A}$, that is, $\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}\left(\mathbf{A}^H\mathbf{A}\right)}$ where $\mathbf{A}^H$ denotes the conjugate transpose of $\mathbf{A}$ [22].

complex-valued downlink channel matrix and $\mathbf{w}_k$ is the kth $N_{TX} \times 1$ normalized linear precoding vector.

The mathematical relationship for the input and output of a precoded MU-MIMO system in the uplink (alternatively known as MAC in the literature) with N_{RX} receive antennas at the base station and one transmit antenna at each user terminal can be written as $\mathbf{y} = \sum_{k=1}^{N_{user}} s_k v_k \mathbf{h}_k + \mathbf{n}$ where $s_k v_k$ is the weighted complex-valued modulated symbol from the kth user, $\mathbf{y} = \left(y_1, y_2, \ldots, y_{N_{RX}}\right)^T$ and $\mathbf{n} = \left(n_1, n_2, \ldots, n_{N_{RX}}\right)^T$ are the $N_{RX} \times 1$ vectors of received signal and noise, respectively, $\mathbf{h}_k$ is the kth $N_{RX} \times 1$ channel vector, where matrix $\mathbf{H} = \left(\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{N_{user}}\right)^T$ is the $N_{RX} \times N_{user}$ complex-valued uplink channel matrix. As mentioned earlier, perfect knowledge of the CSI is necessary at the transmitter in order to achieve the capacity of a multi-user MIMO channel. However, in practical systems, the receiver only provides partial CSI through uplink feedback channels to the transmitter, that is, the multi-user MIMO precoding with limited feedback. The received signal in the downlink of a MU-MIMO system with limited feedback precoding is mathematically expressed as $y_k = \mathbf{h}_k^H \sum_{i=1}^{N_{user}} s_i \hat{\mathbf{w}}_i + n_k, k = 1, 2, \ldots, N_{user}$. The transmit vector for limited feedback precoding is modeled as $\hat{\mathbf{w}}_i = \mathbf{w}_i + \boldsymbol{\varepsilon}_i$ where $\boldsymbol{\varepsilon}_i$ is the error vector generated as a result of the limited feedback and vector quantization, the expression for the received signal can be rewritten as $y_k = \mathbf{h}_k^H \sum_{i=1}^{N_{user}} s_i \mathbf{w}_i + \mathbf{h}_k^H \sum_{i=1}^{N_{user}} s_i \boldsymbol{\varepsilon}_i + n_k, k = 1, 2, \ldots, N_{user}$ where $\mathbf{h}_k^H \sum_{i=1}^{N_{user}} s_i \boldsymbol{\varepsilon}_i$ is the residual interference due to the limited-feedback precoding.

To reduce the residual interference term, one should use more accurate CSI feedback which results in the use of more uplink resources for the feedback. It is shown in the literature that the number of feedback bits per user $N_{feedback}$ must be increased linearly with the SNR γ_{dB} (in decibels) at the rate of $N_{feedback} = (N_{TX} - 1)\log_2\gamma = \gamma_{dB}(N_{TX} - 1)/3$ in order to achieve the full multiplexing gain of N_{TX} antennas [76]. In addition, the scaling of $N_{feedback}$ guarantees that the throughput loss relative to zero-forcing (ZF) precoding with perfect CSI knowledge at the transmitter is upper bounded by N_{TX} bps/Hz, which corresponds to approximately 3 dB power offset. The throughput of a feedback-based ZF system is bounded, if the SNR approaches infinity and the number of feedback bits per user is fixed. Reducing the number of feedback bits according to $N_{feedback} = \alpha\log_2\gamma$ for any $\alpha < N_{TX} - 1$ results in a strictly inferior multiplexing gain of $N_{TX}[\alpha/(N_{TX} - 1)]$ where N_{TX} is the number of transmit antennas and γ is the SNR of the downlink channel.

In order to calculate the amount of feedback required to maintain certain throughput, the difference between the feedback rates of ZF precoding with perfect feedback $R_{PF-ZF}(\gamma)$ and with limited feedback $R_{LF-ZF}(\gamma)$ is required to satisfy the following constraint $\Delta R(\gamma) = R_{PF-ZF}(\gamma) - R_{LF-ZF}(\gamma) \leq \log_2 b$. In order to maintain a rate offset less than $\log_2 b$ (per user) between ZF with perfect CSI and with finite-rate feedback (i.e., $\Delta R(\gamma) \leq \log_2 b, \forall\gamma$), it is sufficient to scale the number of feedback bits per user according to $N_{feedback} = \gamma_{dB}(N_{TX} - 1)/3 - (N_{TX} - 1)\log_2(b - 1)$. The rate offset of $\log_2 b$ (per user) is

translated into a power offset, which is a more useful metric from the design perspective. Since a multiplexing gain of N_{TX} is achieved with ZF, the ZF curve has a slope of N_{TX} bps/Hz/3 dB at asymptotically high SNR. Therefore, a rate offset of $\log_2 b$ bps/Hz per user corresponds to a power offset of $3\log_2 b$ decibels. To feedback $N_{feedback}$ bits through uplink channel, the throughput of the uplink feedback channel should be larger than or equal to $N_{feedback}$, that is, $w_{FB}\log_2(1+\gamma_{FB}) \geq N_{feedback}$ where γ_{FB} denotes the SNR of the feedback channel. Thus the required feedback resource to satisfy the constraint $\Delta R(\gamma) \leq \log_2 b$ can be shown to be given as follows $w_{FB} \geq \left[\gamma_{dB}(N_{TX}-1)/3 - (N_{TX}-1)\log_2(b-1)\right] / \log_2(1+\gamma_{FB})$, that is, the required feedback resource is a function of both downlink and uplink channel conditions [76].

We defined digital precoding as adaptive or non-adaptive weighting of the spatial streams prior to transmission from each antenna port (in a multi-antenna configuration) using a precoding matrix for the purpose of improving the reception or separation of the spatial streams at the receiver. Both feed-back and feed-forward precoder matrix selection schemes can be used in order to select the optimal weights. Feed-back precoding matrix selection techniques do not rely on channel reciprocity, rather they use feedback channels provided that the feedback latency is less than the channel coherence time. In feed-forward approaches, the necessary CSI can be theoretically obtained through direct feedback, where the CSI is explicitly signaled to the transmitter by the receiver or estimated using the SRS. The direct channel feedback methods preclude the channel reciprocity requirement, whereas channel sounding methods rely on channel reciprocity. Therefore, explicit control signaling is required for the PMI-based (feedback) schemes. However, in reciprocity-based schemes, the sounding signals in the uplink and precoded pilots in the downlink are used to assist the transmitter and receiver to appropriately select the precoding matrix. The reciprocity-based schemes have the additional advantage of not being constrained to a finite set of codebooks. Beamforming relies on long-term statistics of the radio channel and, unlike reciprocity-based techniques, does not require short-term correlation between the uplink and downlink in order to properly function.

4.1.9.3.3 Hybrid Beamforming

Hybrid beamforming has been proposed as a possible solution that is able to combine the advantages of both analog and digital beamforming architectures. The idea of hybrid beamforming is based on the concept of phased array antennas commonly used in radar applications. Due to the reduced power consumption, it is also seen as a possible solution for mmWave mobile broadband communication. If the phased array approach is combined with digital beamforming, the phased array approach might also be feasible for non-static or quasi-static scenarios. Considering the inefficiency of mmWave amplifiers and the high insertion loss of RF phase shifters, it is more desirable to perform the phase shifting in the baseband. The power consumption associated with both cases is comparable, as long as the

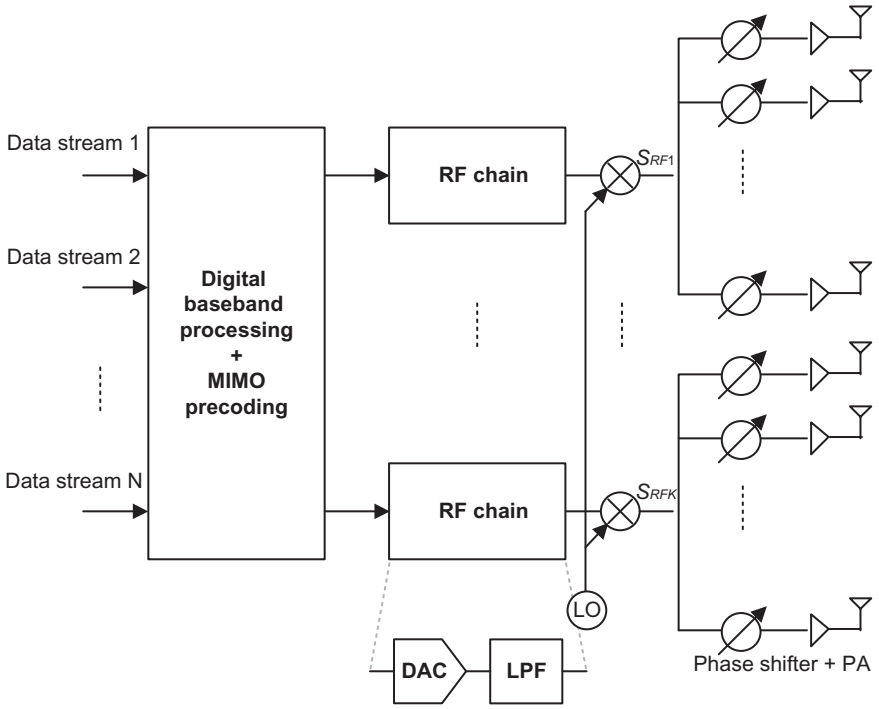


Figure 4.66
Hybrid beamforming architecture.

number of antennas per RF-chain remains relatively small. A significant cost reduction can be achieved by reducing the number of complete RF chains. This does also lead to lower the overall power consumption. Since the number of data converters is significantly lower than the number of antennas, there are less degrees of freedom for digital baseband processing. Thus the number of simultaneously supported streams is reduced compared to digital beamforming. The resulting performance gap is expected to be relatively low in mmWave bands, which this scheme is more suitable, due to the specific channel characteristics. The high-level block diagram of a hybrid beamforming transmitter is shown in Fig. 4.66. The precoding is divided between the analog and digital domains. In theory, it is possible to assume that each amplifier is interconnected to each radiating element.

In recent years, hybrid beamforming with low-resolution data conversion (digital-to-analog/analog-to-digital) has been studied including the energy efficiency/spectral efficiency trade-off of fully connected hybrid and digital beamforming with low-resolution data converters. One of the challenges of large antenna arrays is the increasing cost and complexity of the use of many analog-to-digital and digital-to-analog converters and other RF components to drive individual elements or subarrays. Thus the feasibility study of low-resolution and in

the extreme case one-bit resolution data converters would be very important for practical implementation of massive MIMO systems [16].

In summary, there are three types of beamforming architectures used for antenna arrays [56]:

- Analog beamforming: The traditional way to form beams is to use attenuators and phase shifters as part of the analog RF circuit where a single data stream is divided into separate paths. The advantage of this method is that only one RF chain (PA, LNA, filters, switch/circulator) is required. The disadvantage is the loss from the cascaded phase shifters at high power.
- Digital beamforming: It assumes there is a separate RF chain for each antenna element. The beam is then formed by matrix-type operations in the baseband where the amplitude and phase weighting are applied. For frequencies lower than 6 GHz, this is the preferred method since the RF chain components are comparatively inexpensive and can combine MIMO and beamforming into a single array. For frequencies of 28 GHz and above, the PAs and ADC/DACs are very lossy for standard CMOS components. Gallium arsenide and gallium nitrate can be used in high frequencies to decrease losses at the expense of higher cost.
- Hybrid beamforming: It combines digital and analog beamforming in order to allow the flexibility of MIMO and beamforming while reducing the cost and losses of the beamforming unit. Each data stream has its own separate analog beamforming unit with a set of $N_{antennas}$ antennas. If there are $N_{streams}$ data streams, then there are $N_{streams} \times N_{antennas}$ antennas. The analog beamforming unit loss due to phase shifters can be mitigated by replacing the adaptive phase shifters with a selective beamformer such as a Butler matrix.[21] Some architectures use the digital beamforming unit to steer the direction of the main beam while the analog beamforming unit steers the beam within the digital envelop.

4.1.9.4 Full-Dimension MIMO

Beamforming is a signal processing method that generates directional antenna beam patterns using multiple antennas at the transmitter. It is possible to steer the transmitted signal toward a desired direction and, at the same time, avoid receiving the unwanted signal from an undesired direction. Traditional beamforming schemes controlled the beam pattern only in the horizontal (azimuth) plane. The three-dimensional beamforming adapts the radiation beam pattern in both elevation and azimuth planes to provide more degrees of freedom in supporting users. Higher average user throughput, less inter-cell and inter-sector interference, higher energy efficiency, improved coverage, and increased spectral efficiency are some of the advantages of 3D beamforming or full-dimensional MIMO. In order to exploit the vertical dimension, the antenna tilt can be considered in the vertical axis. The antenna tilt angle is defined as the angle between the horizontal plane and the boresight direction of the antenna pattern. Mechanical alignment of the antenna was traditionally used to adjust

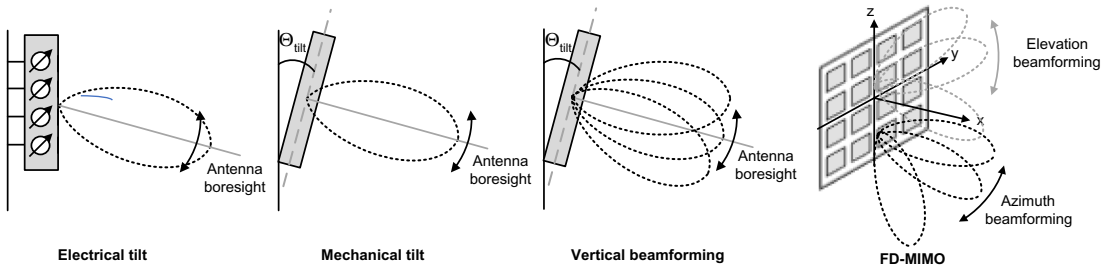


Figure 4.67
Illustration of mechanical/electrical tilting, vertical, and 3D beamforming [23].

the tilt angle of the antenna along the vertical axis. As depicted in Fig. 4.67, some adjustable brackets were used to mechanically change the tilting angle of the antenna.

It is possible to control the tilt angle electrically by applying an overall phase shift to all antenna elements in the array. An active antenna system (AAS) is a recent technology that allows more individual control on antenna elements, where each array element (or group of elements) is integrated with a separate RF transceiver unit that provides remote control to the elements. By employing AAS at the base station, the vertical radiation pattern can also be adjusted dynamically in each sector, and multiple elevation beams can be generated to support multiple users or cover multiple regions; thus full-dimension MIMO (FD-MIMO) is a combination of azimuth and elevation beamforming. Depending on the way that the antenna down tilt is changed, 3D beamforming can be classified into static and dynamic schemes. The static 3D beamforming refers to a system where the antenna tilt at the base station is set to a fixed value according to some statistical metrics, for example, the mean value of the vertical angles of users. This method cannot be adapted to the dynamic patterns of users' movements, that is, once the tilt angle is selected, it will remain unchanged. In contrast the dynamic 3D beamforming is a technique that steers the base station antenna tilt angle according to specific user locations. As mentioned earlier, the antennas at the base station are usually configured as a linear array of a limited number of antennas in the azimuth plane. However, these geometries can shape the radiation pattern only in the horizontal plane; hence, to change the beam in the elevation plane for 3D beamforming, more general 2D or 3D array topologies are necessary. Those arrays are active antenna systems that are spaced in both azimuth and vertical planes with different configurations such as planar, circular, spherical, or cylindrical structures. In addition, the array may include co-polarized or cross-polarized antenna elements [23].

In general, adding more antenna elements to the array provides more flexibility in beam steering designs and increases the number of radiation beams of the array. For vertical sectorization in which the number of vertical sectors is usually small (e.g., two or three), only a small number of antennas are required in the vertical plane. However, in 3D beamforming

with per-user beam pattern adaptation (i.e., user tracking), a large number of antennas are needed. One of the challenges of the 3D beamforming is physical constraints and placement of a large number of antennas at the base station (see Fig. 4.68). This problem may be alleviated in higher frequencies that are used for 5G networks [45].

The study of elevation beamforming and FD-MIMO began in 3GPP LTE Rel-12. In an FD-MIMO system, a base station with two-dimensional active antenna array supports multi-user joint elevation and azimuth or 3D beamforming, which results in much higher cell capacity compared to conventional systems. In an FD-MIMO architecture using 2D AASs with $N_{column} \times N_{row}$ physical antennas, the precoding of a data stream is performed in two stages: (l) antenna-port virtualization that is a stream on an antenna port is precoded on N_{TXRU} transceiver units; and (2) transceiver-unit virtualization where a signal is precoded on $N_{antenna}$ antenna elements. It is noted that in traditional transceiver architecture modeling, a fixed one--to-one mapping is assumed between antenna ports and transceiver units, and TXRU virtualization effect is combined into a fixed antenna pattern which captures the effects of both TXRU virtualization and antenna element pattern. Antenna-port virtualization is an operation in the digital domain, and it refers to digital precoding that can be performed in frequency-selective manner. An antenna port is typically defined in conjunction with a reference signal. For example, for precoded data transmission on an antenna port, a DM-RS is transmitted on the same bandwidth as the data and both are precoded with the same digital precoder. For CSI estimation, on the other hand, CSI-RSs are transmitted on multiple antenna ports. For CSI-RS transmissions, the precoder characterizing the mapping between CSI-RS ports to TXRUs can be designed as an identity matrix, to facilitate device's estimation of TXRU virtualization precoding matrix for data precoding vectors. The TXRU virtualization is an analog operation; thus it refers to time-domain analog precoding. The TXRU virtualization can be made time-adaptive. When TXRU virtualization is semi-static (or rate of change in TXRU virtualization is slow), the TXRU virtualization weights of a serving cell can be chosen to provide good coverage to its serving mobiles and to reduce interference to other cells. There will be more challenges, if TXRU virtualization is dynamic in terms of hardware implementation and protocol design [46].

In 1D TXRU virtualization, N_{TXRU} TXRUs are associated with N_{column} antennas comprising a column antenna array with the same polarization. In 2D antenna arrays with dual-polarized configuration, that is, $P = 2$, and the addition of N_{row} rows, the total number of TXRUs would be $Q = N_{TXRU}N_{row}P$. In 2D TXRU virtualization, Q TXRUs can be associated with any of $N_{column}N_{row}P$ antenna elements. These two different TXRU architectures have different trade-offs in terms of hardware complexity, power efficiency, cost and performance. For each method, subarray partition and full-connection architectures are considered. In subarray partition the antenna elements are partitioned into multiple groups with the same number of elements. In 1D subarray partition, N_{column} antenna elements comprising a column are partitioned into groups of K elements. In 2D subarray partition the total number of antenna elements

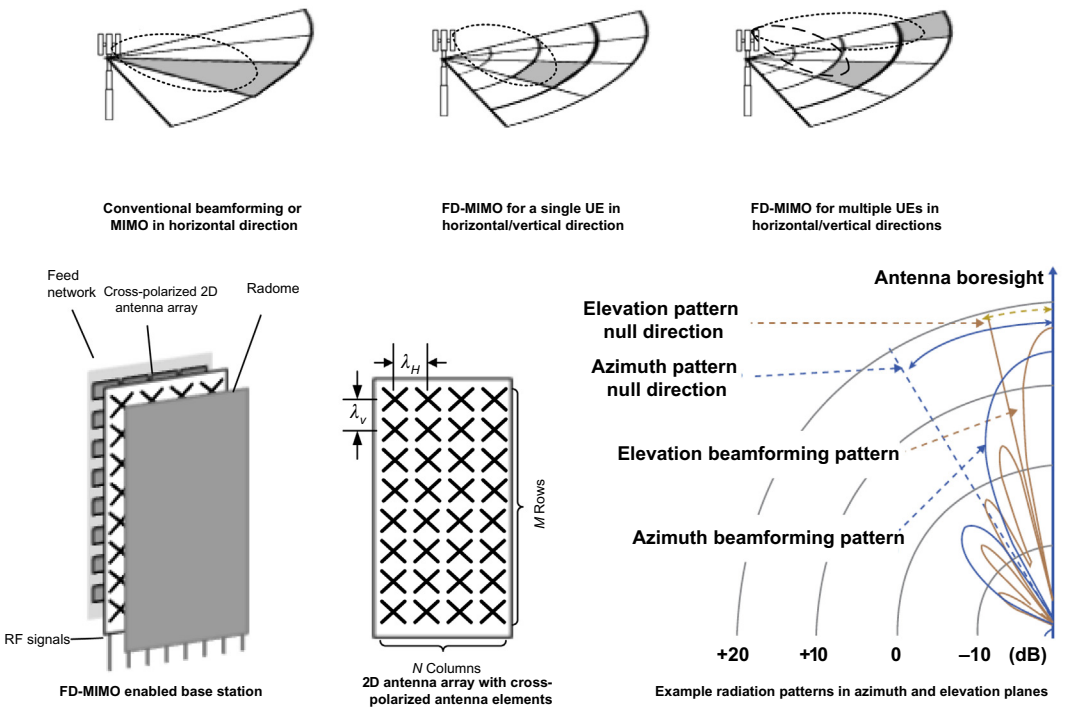


Figure 4.68

Horizontal, vertical, and 3D beamforming FD-MIMO systems: concept of FD-MIMO systems; practical 2D array antenna configuration; vertical and horizontal beamforming patterns; array partitioning architecture with the conventional CSI-RS transmission; and array connected architecture with beamformed CSI-RS transmission [46].

$N_{column}N_{row}P$ is partitioned into rectangular arrays of $K_1 \times K_2$ elements. On the other hand, in 1D full-connection, the output signal of each TXRU associated with a column antenna array with a same polarization is split into N_{column} signals, and those signals are precoded by a group of N_{column} phase shifters or variable gain amplifiers. Then N_{TXRU} weighted signals are combined at each antenna element. In 2D full-connection the output signal of each TXRU is split into $N_{column}N_{row}P$ signals, and those signals are precoded by a group of $N_{column}N_{row}P$ phase shifters or variable gain amplifiers. Then, Q weighted signals are combined at each antenna element. An illustration of 1D subarray partition and full-connection as well as general FD-MIMO architectures are shown in Fig. 4.69 [47].

Multi-antenna systems with a large number of base station antennas, often called massive MIMO, have received much attention in academia and industry as a means to improve the spectral efficiency, energy efficiency, and processing complexity of the cellular systems.

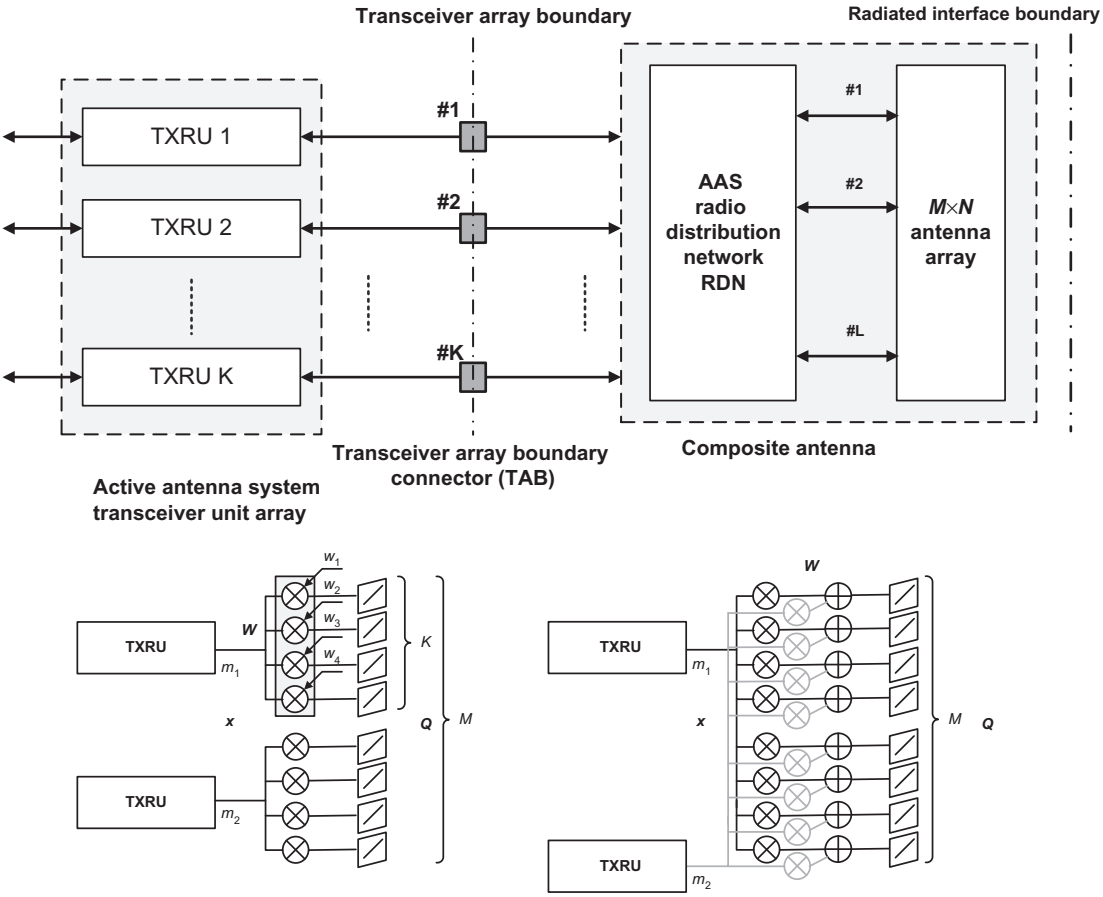


Figure 4.69
Illustration of transceiver architectures in FD-MIMO.

While massive MIMO is a promising technology, there are many practical and technical challenges on the path to its successful commercialization, including design and implementation of low-cost and low-power base station with large antenna arrays, capacity improvement of fronthaul links between remote radio heads and baseband units, measurement and reporting of high dimensional/resolution CSI, etc.

One of the main features of FD-MIMO systems is the potential to use a large number of antennas at the base station. Theoretically, as the number of base station antennas increases, the cross-correlation of two random channel realizations approaches zero; thus inter-user interference in the downlink can be controlled via a simple linear precoder. However, such a benefit can be realized only when the perfect CSI is available at the base station. While the CSI acquisition in TDD systems is relatively simple due to the channel reciprocity, that is not the case for FDD systems, because the time variation and frequency response of the channel in FDD systems are measured via the downlink reference signals and fed back to the base station after quantization. Even in TDD mode, one cannot always rely on channel reciprocity because the measurement at the transmitter does not capture the downlink interference from neighboring cells or co-scheduled UEs. As such, downlink reference signals are still required to capture the CQI for the TDD systems. As a result, downlink reference signals and uplink CSI feedback are crucial for operation of both duplex schemes.

A common problem in closed-loop MIMO systems, and in particular FDD systems, is that the quality of CSI is affected by limitation of the feedback resources. As CSI distortion increases, the MU-MIMO precoder's capability to control the inter-user interference is degraded, resulting in performance degradation of the FD-MIMO system. In general the amount of CSI feedback, which determines the quality of CSI, needs to be scaled with the number of transmit antennas of the base station to control the quantization error, while limiting the overhead of CSI feedback to avoid adverse impact on the system performance. An important problem related to CSI acquisition at the base station is the reference signal overhead. The UE performs channel estimation using the reference signals transmitted from the base station. Since the reference signals are typically distinguished through their orthogonal signatures, their overhead grows linearly with the number of transmit antennas.

As we mentioned earlier, FD-MIMO systems employ 2D planar arrays; thus propagation in both vertical and horizontal directions as well as the geometry of the transmitter array and the propagation effect of the 3D objects between the base station and the mobile station should be taken into account in channel modeling. The 3D channel propagation behavior obtained through measurements show the effect of height and distance-dependent LoS channel and the fact that LoS probability between the base station and the UE increases with the UE's height and increases when the distance between them decreases. Further it shows the effect of height-dependent path loss where the UE experiences less path loss on a higher floor (e.g., 0.6 dB/m gain for a macrocell and 0.3 dB/m gain for a micro cell). The height

and distance-dependent elevation spread of departure (ESD) angles effect is exhibited when the location of the base station is higher than the UE, ESD decreases with the height of the UE and as the UE moves away from the base station [46,47].

FD-MIMO systems make use of a beamformed reference signals for CSI acquisition. Beamformed reference signal transmission is a channel training technique that uses multiple precoding weights in the spatial domain. In this scheme, the UE selects the best weight among those transmitted and then feeds back this index. This scheme provides many benefits compared to the case with non-precoded reference signals and especially when the number of transmit antennas is large. It can be shown that this scheme has less uplink feedback overhead relative to the case with perfect CSI, where the number of feedback bits used for channel vector quantization is linearly proportional to the number of transmit antennas, whereas the amount of feedback for the beamformed reference signals scales logarithmically with the number of reference signals, because the UE only feeds back the index of the best beamformed reference signal. It can be further shown that there is less downlink pilot overhead when the non-precoded reference signal is used. The non-beamformed reference signal overhead increases with the number of transmit antennas, resulting in substantial loss of sum capacity in the FD-MIMO, whereas the beamformed reference signal overhead is proportional to the number of reference signals and independent of the number of transmit antennas; therefore, the rate loss of the beamformed reference signals is marginal even when the number of transmit antennas increases [46,47].

As we mentioned earlier, an AAS transceiver contains integrated PA and LNA so that the gNB can control the gain and phase of individual antenna elements. A radio signal distribution/combining network between TXRUs and antenna elements was introduced (see Fig. 4.69) whose role is to deliver the transmit signal from the PA to the antenna array elements and the received signal from the antenna array to the LNA. Depending on the CSI-RS transmission and feedback mechanism, two architecture options, array partitioning and array connected, may be used. The former architecture is more suitable for the conventional codebook scheme, and the latter is for the beamforming scheme. In the array partitioning architecture, antenna elements are divided into multiple groups, and each TXRU is connected to one of them, whereas in the array connected architecture, the radio distribution network is designed such that the RF signals of multiple TXRUs are delivered to the single-antenna element. To combine RF signals from multiple TXRUs, additional RF combining circuitry is needed. In the array partitioning architecture, the total number of antenna elements L is partitioned into several groups of TXRUs, and an orthogonal CSI-RS is assigned for each group. Each TXRU transmits its own CSI-RS so that the UE can measure channel $\mathbf{h}$ from the CSI-RS observation. In the array connected architecture, each antenna element is connected to $L' < L$ TXRUs and an orthogonal CSI-RS is assigned for each TXRU. Denoting $\mathbf{h} \in \mathbb{C}^{1 \times N}$ as the channel vector and $\mathbf{v} \in \mathbb{C}^{N \times 1}$ as the precoding vector for each beamformed CSI-RS, the beamformed CSI-RS observation can be expressed as $y = \mathbf{h}\mathbf{v}x + n$ and the UE measures the precoded channel $\mathbf{h}\mathbf{v}$. Due to the narrow and directional

CSI-RS beam transmission with a linear array, the SNR of the precoded channel is maximized at the target direction, that is, $SNR = \left|\mathbf{hv}(\varphi)\right|^2/\sigma^2$ where φ is the beam direction and σ^2 is the noise variance. In non-beamformed scenario, the UE selects and sends a precoder index which maximizes certain performance criterion to the gNB and adapts to the channel variation. In the beamformed scenario, the gNB transmits multiple beamformed CSI-RSs using the connected array architecture and the UE selects the preferred beam and then feeds back its index. When the gNB receives the beam index, the weight corresponding to the selected beam is used for data transmission to the UE [48].

Let us consider a cellular system consisting of N_{cell} cells each with one base station and N_{UE} terminals in each cell, as shown in Fig. 4.70. Each gNB is equipped with a 2D antenna array of $N_V \times N_H$ vertical and horizontal antennas, and each UE has a single antenna. We assume that all gNBs and UEs are synchronized and operate in TDD mode with universal frequency reuse. In the downlink, the lth base station applies a $N_V N_H \times N_{UE}$ precoder $\mathbf{F}_n, n = 1, 2, \ldots, N_{cell}$ to transmit a symbol for each user, with a power constraint $\left\|[\mathbf{F}_n]_{:,k}\right\|^2 = 1, k = 1, 2, \ldots, N_{UE}$. Uplink and downlink channels are assumed to be reciprocal. If $\mathbf{h}_{nck}$ denotes the $N_V N_H \times 1$ uplink channel from user k in cell c to the nth base station, then the received signal by this user in the downlink can be

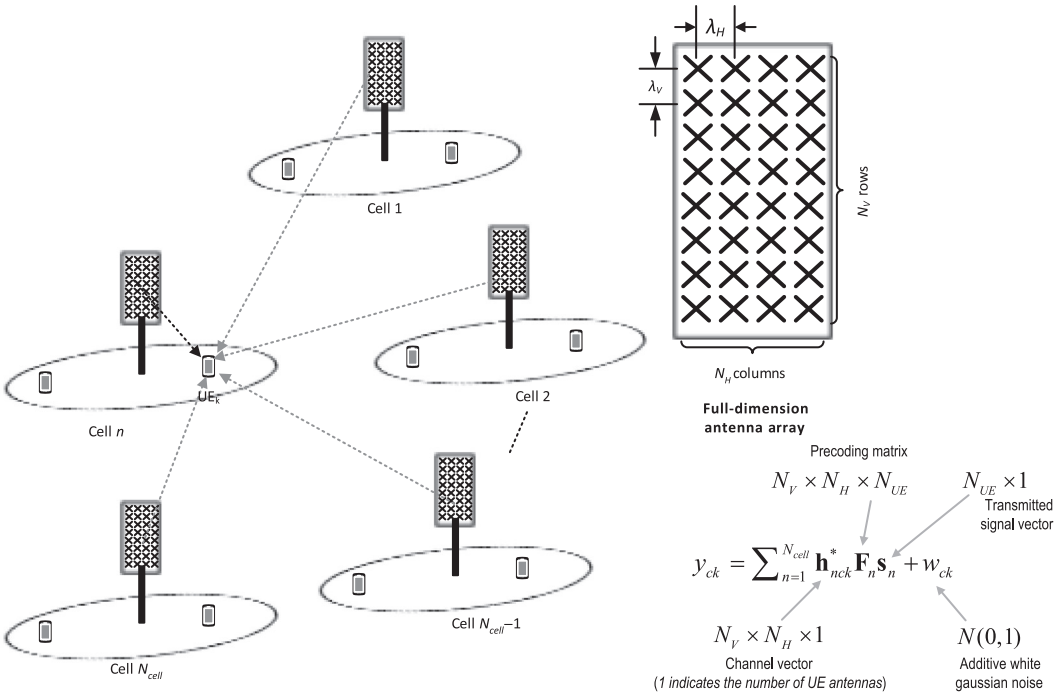


Figure 4.70
Illustration of full-dimension MU-MIMO concept [48].

written as $y_{ck} = \sum_{n=1}^{N_{cell}} \mathbf{h}_{nck}^{*} \mathbf{F}_n \mathbf{s}_n + w_{ck}$ where $\mathbf{s}_n$ represents the $N_{UE} \times 1$ vector of transmitted symbols from the nth base station. We further assume $E\{\mathbf{s}_n \mathbf{s}_n^{*}\} = P/N_{UE}\mathbf{I}$ with P denoting the total transmit power and $w_{ck} \sim N(0, \sigma^2)$ is the additive white Gaussian noise at kth user receiver in cell c. Given the 2D antenna arrays deployed at the gNBs, the channels from the base stations to each UE have a 3D structure. Using the Kronecker product correlation model, which has been shown to provide a good approximation to 3D covariance matrices, the covariance of the 3D channel $\mathbf{h}_{nck}$, which is defined as $\mathbf{R}_{nck} = E\{\mathbf{h}_{nck}\mathbf{h}_{nck}^{*}\}$, is approximated by $\mathbf{R}_{nck} = \mathbf{R}_{nck}^{\mathbf{A}} \otimes \mathbf{R}_{nck}^{\mathbf{E}}$ where $\mathbf{R}_{nck}^{\mathbf{A}}$ and $\mathbf{R}_{nck}^{\mathbf{E}}$ represent the covariance matrices in the azimuth and elevation directions, respectively. If $\mathbf{R}_{nck}^{\mathbf{A}} = \mathbf{U}_{nck}^{\mathbf{A}} \mathbf{\Lambda}_{nck}^{\mathbf{A}} \mathbf{U}_{nck}^{\mathbf{A}^*}$ and $\mathbf{R}_{nck}^{\mathbf{E}} = \mathbf{U}_{nck}^{\mathbf{E}} \mathbf{\Lambda}_{nck}^{\mathbf{E}} \mathbf{U}_{nck}^{\mathbf{E}^*}$ are the SVDs of $\mathbf{R}_{nck}^{\mathbf{A}}$ and $\mathbf{R}_{nck}^{\mathbf{E}}$ then using Karhunen−Loève transformation[17], the channel $\mathbf{h}_{nck}$ can be expressed as

$\mathbf{h}_{nck} = [\mathbf{U}_{nck}^{\mathbf{A}} \mathbf{\Lambda}_{nck}^{\mathbf{A}\ 1/2} \otimes \mathbf{U}_{nck}^{\mathbf{E}} \mathbf{\Lambda}_{nck}^{\mathbf{E}\ 1/2}]\mathbf{w}_{nck}$ where $\mathbf{w}_{nck} \sim N(0, \mathbf{I})$ is a $\text{rank}(\mathbf{R}_{nck}^{\mathbf{A}})\text{rank}(\mathbf{R}_{nck}^{\mathbf{E}}) \times 1$ vector, with $\text{rank}(\mathbf{A})$ representing the rank of the matrix $\mathbf{A}$ [48]. The SINR at the kth user receiver in cell c can be shown to be as follows:

$$\text{SINR}_{ck} = \frac{\overbrace{(P/N_{UE})\left|\mathbf{h}_{cck}^{*}[\mathbf{F}_c]_{:,k}\right|^2}^{\text{Received signal power of the } k\text{th UE}}}{\underbrace{(P/N_{UE})\sum_{m \neq k}\left|\mathbf{h}_{cck}^{*}[\mathbf{F}_c]_{:,m}\right|^2}_{\substack{\text{Sum of interference from the signal}\\ \text{transmited to all other UEs in cell c}\\ \text{(interference)}}} + \underbrace{(P/N_{UE})\sum_{n \neq c}\left\|\mathbf{h}_{nck}^{*}\mathbf{F}_n\right\|^2}_{\substack{\text{Sum of interference from the signal}\\ \text{transmitted by all neighboring cells}\\ \text{(interference)}}} + \underbrace{\sigma^2}_{\text{Receiver noise}}}$$

The objective is to design the precoding matrices $\mathbf{F}_n, n = 1, 2, \ldots, N_{cell}$ such that they minimize inter-cell interference with minimal requirements on the channel knowledge, and they

[17] In the theory of stochastic processes, the Karhunen−Loève theorem is a representation of a stochastic process as an infinite linear combination of orthogonal functions. The transformation is also known as Hotelling transform and eigenvector transform, and it is closely related to principal component analysis (PCA). In contrast to a Fourier series where the coefficients are fixed numbers and the basis functions are sinusoidal the coefficients in the Karhunen−Loève theorem are random variables and the basis functions depend on the process. In fact the orthogonal basis functions used in this representation are determined by the covariance matrix of the process. Therefore, the Karhunen−Loève transform adapts to the process in order to produce the optimal basis for its expansion. In the case of a centered stochastic process $\{X(t)|t \in [a, b]\}$, that is, $E[X(t)] = 0 \forall t \in [a, b]$, satisfying a continuity condition, it can be shown that $X(t)$ can be expanded as $X(t) = \sum_{k=1}^{\infty} Z_k e_k(t)$ where Z_k's are pairwise uncorrelated random variables and the functions $e_k(t)$ are continuous real-valued functions on $[a, b]$ that are pairwise orthogonal in $L^2[a, b]$. It is therefore sometimes said that the expansion is bi-orthogonal since the random coefficients Z_k are orthogonal in the probability space while the deterministic functions $e_k(t)$ are orthogonal in the time domain. The general case of a process $X(t)$ that is not centered can be converted into a centered process by considering $X(t) - E[X(t)]$, which is a centered process. If the process is Gaussian then the random variables Z_k are Gaussian and stochastically independent. This result generalizes the Karhunen−Loève transform. An important example of a centered real stochastic process on $[0, 1]$ is the Wiener process, where the Karhunen−Loève theorem can be used to provide a canonical orthogonal representation for it. In this case, the expansion consists of sinusoidal functions [22].

can be implemented using low-complexity hybrid analog/digital architectures, that is, with a small number of RF chains. The main idea of multi-layer precoding is to design the precoder matrix as a product of three precoding matrices (layers) where each layer is designed to achieve only one precoding objective, for example, maximizing desired signal power, minimizing inter-cell interference, or minimizing multi-user interference [48].

$$\mathbf{F}_c = \underbrace{\mathbf{F}_c^{(1)}}_{\substack{\text{Inter}-\text{cell interference management} \\ \downarrow \\ \text{Minimize inter}-\text{cell interference}}} \quad \underbrace{\mathbf{F}_c^{(2)}}_{\substack{\text{Desired signal} \\ \text{beamforming}}} \quad \underbrace{\mathbf{F}_c^{(3)}}_{\substack{\text{Multi}-\text{user interference management} \\ \downarrow \\ \text{Minimizing multi}-\text{user interference}}}$$

4.1.9.5 Large-Scale (Massive) MIMO Systems

Massive MIMO is the generalization of a multi-user MIMO system that serves multiple users through spatial multiplexing over a channel with favorable propagation[18] conditions using time-division duplex scheme and relying on channel reciprocity and uplink reference signals to obtain CSI of each user. The base station is equipped with $N_{antennas}$ antennas to communicate with N_{user} (typically modeled with single-antenna) UEs on each time/frequency resource, where $N_{user} \ll N_{antennas}$. Each base station in the network operates individually and processes its signals using linear transmit precoding and linear receive combining [16,52]. By coherent processing of the signals over the array, transmit precoding can be used in the downlink to focus each signal at its target user, and receive combining can be used in the uplink to distinguish between signals received from different user terminals, thus the larger the number of antennas, the finer the spatial precision. A generic massive MIMO system operates in TDD mode, where the uplink and downlink transmissions take place on the same frequency resource but are separated in time. The physical propagation channels are reciprocal, meaning that the channel responses are theoretically the same in both directions, which can be utilized in TDD operation. In practice, the transceiver hardware is not reciprocal, thus transceiver calibration is required to exploit the channel reciprocity. Since uplink–downlink hardware mismatches only slowly and slightly change over time, they can be mitigated by simple calibration methods even without extra reference transceivers by relying on mutual coupling between antennas in the array. There are several reasons for the suitability of the TDD mode for massive MIMO which include the following [16,52]:

- The base station needs to know the CSI to process the antennas coherently.

[18] Favorable propagation means that the channel matrix between the base station antenna array and the users is well-conditioned. In a massive MIMO system, under some conditions, the favorable propagation property holds due to the law of large numbers. In other words, the propagation is said to be favorable when users are mutually orthogonal in some practical sense.

- The uplink channel estimation overhead is proportional to the number of terminals and independent of the number of antennas, making the scheme scalable with respect to the number of antennas. Furthermore, basic estimation theory indicates that the estimation quality (per antenna) cannot be reduced by adding more antennas at the base station. In fact the estimation quality improves with the number of antennas, if there is a known correlation structure between the channel responses over the array.

The data transmission in massive MIMO is based on linear processing at the gNB. In the uplink, the gNB has N_{RX} observations of the multiple access channel from the N_{user} terminals. The gNB applies maximal ratio combining to separate the signal transmitted by each terminal from the interfering signals, using the channel estimate of a terminal to maximize the signal power of that terminal by coherently adding the signal components. This results in a signal amplification proportional to N_{RX}, which is known as the array gain. Alternatively, ZF combining can be used, which suppresses inter-cell interference at the cost of reducing the array gain to $N_{RX} - N_{user} + 1$, or MMSE combining can be utilized that balances between amplifying signals and suppressing interference. Receive combining creates one effective scalar channel per terminal where the intended signal is amplified, and/or the interference is suppressed. The performance of the received combining methods will be improved by adding more gNB antennas, since there are more channel observations to utilize. The remaining interference is typically treated as additive noise; thus, conventional single-user detection algorithms can be applied. Another benefit of the combining is that small-scale fading averages over the array, in the sense that its variance decreases with N_{RX}. This is known as channel hardening and is a consequence of the law of large numbers. Since the uplink and downlink channels are ideally reciprocal in TDD systems, there is a strong connection between receive combining in the uplink and transmit precoding in the downlink. This is known as uplink–downlink duality. Linear precoding based on MRC, ZF, or MMSE principles can be applied to focus each signal on its target user and possibly to minimize interference toward other users [16,52].

It can be shown that the achievable spectral efficiency per cell of massive MIMO systems under ideal conditions and independent identically distributed Rayleigh fading can be expressed in the following form [16,52]:

$$\eta = N_{user}\left(1 - \frac{N_{user}}{\tau}\right)\log_2\left(1 + \frac{\varepsilon_{CSI}\gamma N_{TX}}{N_{user}\gamma + 1}\right) \quad \text{bps/Hz/Cell}, \quad \varepsilon_{CSI} = \left(1 + \frac{1}{N_{user}\gamma_{uplink}}\right)^{-1}$$

where $\left(1 - N_{user}/\tau\right)$ is the loss due to pilot transmission, γ is the downlink/uplink SNR, and ε_{CSI} is the quality of the estimated CSI, proportional to the mean-squared power of the MMSE channel estimate, where $\varepsilon_{CSI} = 1$ represents perfect CSI. Note that the numerator of the logarithm argument increases proportionally with respect to N_{TX} due to the array gain

and that the denominator represents the interference plus noise. While the generic theory of massive MIMO systems assumed single-antenna terminals, the technology can support terminals with N'_{RX} antennas. In this case, N_{user} denotes the number of simultaneous data streams and the preceding equation describes the spectral efficiency per stream. These streams can be divided over anything from N_{user}/N'_{RX} to N_{user} terminals [16,52].

We discussed the capacity of MIMO systems earlier in this chapter, which can be written as follows:

$$C = \log_2 \det\left(\mathbf{I} + \frac{\gamma}{N_{TX}}\mathbf{H}\mathbf{H}^H\right) = \sum_{i=1}^{\min(N_{TX},N_{RX})} \log_2\left(1 + \frac{\gamma\sigma_i^2}{N_{TX}}\right)$$

In the preceding equation, it is assumed that transmitter has the full knowledge of CSI and that the channel matrix $\mathbf{H}$ can be decomposed using SVD method, where $\sigma_i^{2\prime}$s are the eigenvalues of $\mathbf{H}\mathbf{H}^H$. In the preceding equation, if the SNR γ is extremely small, the capacity asymptotically approaches $C_{\gamma\to 0} \approx \gamma N_{RX}/\ln 2$, which is independent of N_{TX}, thus, even under the most favorable propagation conditions, the multiplexing gains are lost, and from the perspective of achievable rate, multiple transmit antennas are of no value. Next, let the number of transmit antennas grow large while keeping the number of receive antennas constant. We further assume that the row vectors of the channel matrix are asymptotically orthogonal, $C_{N_{TX}\gg N_{RX}} \approx N_{RX}\log_2(1+\gamma)$. Then, let the number of receive antennas increase while keeping the number of transmit antennas constant. We also assume that the column vectors of the channel matrix are asymptotically orthogonal, $C_{N_{RX}\gg N_{TX}} \approx N_{TX}\log_2\left(1+\gamma N_{RX}/N_{TX}\right)$. Therefore, an excess number of transmit or receive antennas, combined with asymptotic orthogonality of the propagation vectors, constitutes a highly desirable scenario. Additional receive antennas continue to improve the effective SNR and could in theory compensate for a low SNR and restore multiplexing gains that would otherwise be lost. Furthermore, orthogonality of the propagation vectors implies that independent and identically distributed complex-Gaussian inputs are optimal so that the achievable rates are in fact the true channel capacities [61].

The studies on massive MIMO have been mainly focused on frequencies below 6 GHz, where the transceiver hardware technology is very mature. The same concept can be applied in mmWave bands, where many antennas might be required since the effective aperture of the antenna is much smaller. However, the hardware implementation will be more challenging. The support of mobility will be more difficult because the coherence time will be an order of magnitude shorter due to higher Doppler spread, which reduces the spatial multiplexing capability.

The channel impulse response between a user terminal and a base station can be represented by an $N_{antennas}$-dimensional vector. Since the N_{user} channel vectors are mutually nonorthogonal in general, advanced interference cancellation receivers are needed to suppress

interference and achieve the sum capacity of the multi-user channel. As we mentioned earlier, the favorable propagation is an environment where the N_{user} users' channel vectors are mutually orthogonal (i.e., their inner products are zero). The favorable propagation channels are ideal for multi-user transmission since the interference is removed by simple linear processing (i.e., MRC and ZF) that utilizes the channel orthogonality. The question is whether there are any favorable propagation channels in practice. An approximate form of favorable propagation is achieved in non-LOS environments with rich scattering, where each channel vector has independent stochastic entries with zero mean and identical distribution. Under these conditions, the inner products (normalized by $N_{antennas}$) approach to zero as the number of antennas increases; meaning that the channel vectors tend to be orthogonal as $N_{antennas}$ increases. The sufficient condition above is satisfied for Rayleigh fading channels, which are considered in the studies on massive MIMO, but approximate favorable propagation can also be obtained in other conditions [16,52].

The conventional open-loop beamforming provides meaningful array gains for small arrays in LoS propagation environments; however, this scheme is not scalable and not able to handle isotropic fading (*isotropic fading encompasses a broad range of fading channels with the common property that the transmitter is unable to track the directions of the users' time-varying channel vectors. Thus, from the transmitter standpoint all directions are statistically equivalent*). In practice, the channel of a particular user terminal might not be isotropically distributed, rather it might have distinct statistical spatial properties. The codebook in open-loop beamforming cannot be tailored to a specific terminal, rather needs to explore all channel directions that are possible for the array. For large arrays with arbitrary propagation properties, the channels must be measured by reference signals as is done in the massive MIMO [16,52].

The studies on massive MIMO were mainly focused on the asymptotic regime where the number of service antennas $N_{antennas} \to \infty$. Recent studies have derived closed-form achievable spectral efficiency expressions that are valid for any number of antennas and user terminals, SNR, and choice of reference signals. Those expressions do not rely on idealized assumptions such as perfect CSI, rather on worst case assumptions regarding the channel acquisition and signal processing. Although the total spectral efficiency per cell is greatly improved with massive MIMO, the anticipated performance per user lies in the conventional range of $1-4$ bps/Hz. This is part of the range where conventional channel codes perform close to the Shannon limits.

There are no strict requirements on the relation between $N_{antennas}$ and N_{user} in massive MIMO systems. A simple definition of massive MIMO would be a system with many active antenna elements that can serve a large number of user terminals. One should avoid specifying a certain ratio $N_{antennas}/N_{user}$, since it depends on a variety of conditions such as the system performance metric, propagation environment, and coherence block length.

The massive MIMO gains do not require high-precision hardware; in fact, lower hardware precision can be handled compared to other systems since additive distortions are suppressed in the processing. Another reason for the robustness is that massive MIMO can achieve high spectral efficiencies by transmitting low-order modulations to a multitude of terminals, while contemporary systems require high-precision hardware to support transmission of high-order modulations to a few terminals [16,52].

In an OFDM system, resource allocation means that the time-frequency resources are divided between the terminals to satisfy user-specific performance constraints, finding the best subcarriers for each terminal, and overcoming the small-scale fading effects by power control. Frequency-selective resource allocation can provide significant improvements when there are large variations in channel quality over the subcarriers, but it is also demanding in terms of channel estimation and computational overhead since the decisions depend on the small-scale fading, which varies in time. If the same resource allocation concepts were applied to massive MIMO systems, with tens of terminals at each of the thousands of subcarriers, the system complexity would have been excessively prohibitive. However, the channel hardening effect in massive MIMO means that the channel variations are negligible over the frequency-domain and mainly depend on large-scale fading in the time domain, which typically varies much slower than small-scale fading, making the conventional resource allocation concepts unnecessary for massive MIMO. The available bandwidth can be simultaneously allocated to each active terminal, and the power control decisions are made jointly for all subcarriers based only on the large-scale fading characteristics [16,52]. Thus, the resource allocation can be greatly simplified in massive MIMO systems.

4.1.9.6 NR Multi-antenna Transmission Schemes

Multi-antenna transmission and beamforming of control and traffic channels are the distinct features of the new radio relative to its predecessors. In above 6 GHz frequency bands, the large number of antenna elements is primarily used for beamforming to achieve enhanced coverage, while at lower frequency bands, they enable FD-MIMO and interference avoidance by spatial filtering. The NR physical channels and signals, including those used for control and synchronization, have all been designed for beamforming. Unlike LTE whose downlink control channels used transmit diversity to ensure sufficient control channel link budget, NR control channels rely on a single-antenna port and beamforming to achieve coverage requirements. The CSI for operation of massive MIMO schemes can be obtained by feedback of CSI reports based on transmission of CSI reference signals in the downlink, either per antenna element or per beam, as well as using uplink measurements exploiting channel reciprocity. In order to simplify the implementation, the new radio supports analog beamforming in addition to digital precoding and beamforming. The support of analog beamforming, where the beam is shaped after digital-to-analog conversion, is necessary in high frequencies. Analog beamforming requires that the receive or transmit beams to be

formed in one direction at a given time and further requires beam-sweeping, in which the same signal is repeated over multiple OFDM symbols and in different transmit beams. This is to ensure that control/traffic signals can be transmitted with high gain to sufficiently cover the service area of the base station. Beam management procedures and signaling are further specified in NR including indication to the device to assist selection of the receive beam (in case of analog receive beamforming) during data and control reception [11].

The beams corresponding to the large antenna arrays are narrower and beam tracking may fail; therefore, beam recovery procedures have been specified in NR where a device can trigger a beam recovery procedure. Furthermore, a cell may consist of multiple transmission points, each transmitting its own beams; in that case, beam management procedures would allow device-transparent mobility for seamless handover between the beams of different transmission points. In addition, uplink-centric and reciprocity-based beam management is possible by utilizing uplink reference signals. The possibility of spatially separating users increases when using a large number of antenna elements in lower frequency bands in both uplink and downlink; however, this requires the knowledge of channel at the transmitter. In NR, extended support for such multi-user spatial multiplexing was introduced, either through high-resolution CSI feedback with a linear combination of DFT vectors, or uplink SRS improvements based on channel reciprocity. Moreover, support for distributed MIMO has been introduced, where the device can receive multiple PDCCHs and PDSCHs per slot to enable simultaneous data transmission from multiple transmission points to the same user [69].

The support of hybrid beamforming and high-resolution CSI feedback have been two important design principles in NR MIMO. The first goal is addressed by beam management, where a UE measures a set of analog beams for each digital port and reports the beam quality. The gNB then assigns one or a small number of analog beams to the UE. As the downlink channel experienced by the UE changes, the gNB can modify this assignment, particularly when the link associated with the assigned beam fails. While beam management is instrumental in above 6 GHz frequency bands, it is also applicable to sub-6 GHz bands. For instance, in multipoint transmission scenario, where multiple transmission-reception points are associated with a UE, each link corresponds to a beam. The second goal is addressed by designing a modular and scalable CSI framework. High-resolution spatial channel information is provided via two-stage high-resolution precoding. The first stage involves the choice of basis subset, and the second stage comprises a set of coefficients for approximating a channel eigenvector with a linear combination of the basis subset. Note that while beam management and CSI acquisition can be operated independently, they can be used together for mobile UEs.

In NR, the DM-RS and the TRS are used to estimate the timing offset and frequency errors. To guarantee that the timing-offset and frequency-error of the antenna panels can be estimated independently, the DM-RS and TRS are grouped, and the grouping information is signaled to the UE. The UE distinguishes the time-frequency resource of reference signals

allocated for different data streams. In this way, timing-offset and frequency-errors can be independently estimated for different MIMO layers in non-coherent MIMO transmission to improve the performance.

For downlink transmission with multiple antenna panels, the accuracy of CSI acquisition is critical for the system performance. To characterize the channel directional information, the codebook and the related feedback mechanism are typically designed for CSI acquisition, especially for FDD systems. For users in the cell edge of TDD systems, reciprocal channel estimation based on uplink reference signals is also inaccurate, and codebook-based CSI feedback can help improve the performance. Codebook design for massive MIMO should be flexible for different antenna array structures and applicable to various numbers of antennas. For a uniform panel array, the codebook is designed similar to the one used for FD-MIMO. While a single DFT vector can only characterize one spatial channel path, an advanced codebook taking the combination of two or more DFT vectors as the precoder can capture the characteristics of multipath channels. For non-uniform panel array, the antenna elements in the horizontal/vertical direction cannot be viewed as a uniform array. Thus array response vector is not in DFT form, and additional phase difference exists between the panels. In other words, the DFT vector cannot capture the actual channel response but can cause beam distortion and beam gain reduction. Moreover, panels at one transmit/receive point are not easily calibrated in a practical implementation, where a fixed or random phase may exist among different panels. A good codebook design should use phase/amplitude factors among panels to combine the beamforming vectors of the panels to match the array response [59].

In multi-panel MIMO codebook design, 2D-DFT vectors may be used as the per-panel beamforming vectors representing the spatial propagation properties for each panel. Furthermore, an inter-panel co-phasing factor is added across the DFT-based codeword to reduce performance loss. Let us use the two-panel case as an example, one can design the multi-panel codebook $\mathbf{W}$ for dual-polarized antenna array as $\mathbf{W} = \begin{bmatrix} \mathbf{W}_1^1\mathbf{W}_2^1 & \mathbf{W}_1^2\mathbf{W}_2^2\varphi \end{bmatrix}^T$ where $\mathbf{W}_1^1$ and $\mathbf{W}_2^1$ consist of DFT vectors reflecting long-term channel characteristics of antenna panel 1 and panel 2, respectively; $\mathbf{W}_1^2$ and $\mathbf{W}_2^2$ consist of phase factors reflecting the short-term and frequency-selective channel elements of panels 1 and 2, respectively, and φ is the inter-panel co-phasing factor, which is designed for feedback in a wideband manner or subband manner according to the use cases. The precoding vectors for both panels are restricted to be the same in order to reduce the feedback overhead [59].

For a uniform or non-uniform planer array with ideal synchronization or small phase offset, coherent transmission can be used to achieve spatial multiplexing/diversity gain. However, the performance of coherent transmission may be degraded in practice due to different reflection and refraction propagation paths which cause different average channel delay and different average channel gain for a given direction of arrival and departure. If antenna ports from different panels experience different large-scale fading, it causes the eigenvectors of

the channel matrix to become ill-conditioned (rank deficiency), resulting in inaccurate CSI for coherent MIMO transmission. Moreover, time-varying amplitude or phase calibration error among the panels may occur in practice, especially when the antenna panels have independent clocks and different operating temperatures. In addition, the frequency offset for different panels observed at the receiver may be different when there is a time-varying relative phase between different antenna panels.

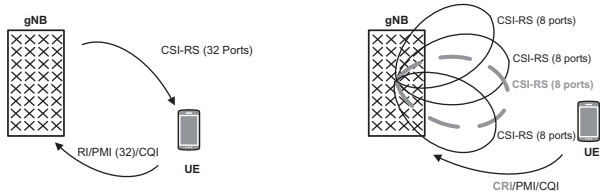
The reference signals transmitted from geographically separated panels may experience different Doppler shift and Doppler spread, because the UE may move in a different directions relative to the panels. If the UE assumes the panels have the same Doppler shift and Doppler spread, the frequency offset estimation would be inaccurate, which results in performance degradation. The relative frequency offset between panels can be from 0 to 300 Hz, due to the following factors [59]:

- Doppler shift: For a UE moving with the speed of 30 km/h at the carrier frequency of 2 GHz, the maximum difference between the experienced Doppler shifts is 111 Hz.
- Frequency error: In LTE, the maximum tolerable oscillator inaccuracy is ± 0.05 ppm for the wide area base station classes; thus the maximum frequency error between two non-calibrated panels with independent oscillators at the carrier frequency of 2 GHz is about 200 Hz. Thus the total amount of relative frequency offset is around 300 Hz. As the carrier frequency increases, the frequency offset problem would be more severe.

The same timing offset assumption for different panels may impact the accuracy of channel estimation. With a timing offset, a random linear phase factor will be added to the channel coefficient on adjacent subcarriers, which is difficult to accurately estimate. The inter-symbol interference caused by the timing offset impacts the channel estimation accuracy in both CSI measurement and data demodulation, resulting in significant MIMO performance degradation. It can be shown that the performance loss due to timing offset is not negligible, especially in the case of negative timing offset. For positive timing offset, when the aggregated signals are received with the timing difference shorter than the cyclic prefix length, less ISI is incurred. For the negative timing offset, since the incurred ISI cannot be mitigated by removing the cyclic prefix, severe performance loss is caused. The capability of interference suppression in the receiver is critical for the performance of non-coherent MIMO transmission. For coherent MIMO transmission, a traditional linear receiver such as MRC/MMSE can be used to achieve acceptable performance, while for non-coherent MIMO transmission, an advanced receiver is required. If the data streams from different directions (especially in the case of widely spaced panels) arrive at the UE in different angles, a simple linear MMSE receiver may be sufficient. However, the performance gain depends on the remaining interference after the space-domain interference rejection by the MMSE processing. If different streams arrive at the UE from the same or adjacent directions (e.g., different antenna panels located in a centralized manner),

Downlink MIMO operation in sub-6 GHz

Single CSI-RS	Multiple CSI-RS	SRS-based
▪ CSI-RS may be beamformed ▪ Allows codebook feedback ▪ Similar to LTE class-A CSI feedback (gNB transmit CSI-RS; UE computes RI/PMI/CQI) ▪ Maximum of 32 ports in the CSI-RS (codebooks are defined for up to 32 ports) ▪ Typically intended for arrays having 32 TXRUs or less with no beam selection (no CRI)	▪ Combines beam selection with codebook feedback (multiple beamformed CSI-RS with CRI feedback) ▪ Similar to LTE Class-B CSI Feedback (gNB transmits one or more CSI-RS, each in different directions; UE computes CRI/PMI/CQI) ▪ Supports arrays with arbitrary number of TXRUs ▪ Maximum of 32 ports per CSI-RS	▪ Exploits TDD reciprocity feature ▪ Similar to SRS-based operation in LTE ▪ Supports arrays with arbitrary number of TXRUs ▪ Procedure (UE transmits SRS, gNB computes precoding weights)



Downlink MIMO operation in above 6 GHz

Single-panel array	Multipanel array
▪ Combination of analog beamforming and digital precoding at baseband ▪ Analog beamforming is typically one RF beamforming weight-vector per polarization (a single cross-polarized beam) ▪ Supports two TXRUs and single-user MIMO only ▪ Digital precoding options: None (rank-2 all the time); CSI-RS based (RI/PMI/CQI); and SRS-based (RI/CQI)	▪ Combination of analog beamforming and digital precoding at baseband ▪ Analog beamforming is typically one RF beamforming weight-vector per polarization per panel ▪ One cross-polarized beam per subpanel ▪ Number of TXRUs = 2 x number of panels ▪ Digital precoding options: CSI-RS based (RI/PMI/CQI); SRS-based (RI/CQI); and SU-MIMO and MU-MIMO (typically one UE per cross-polarized beam)

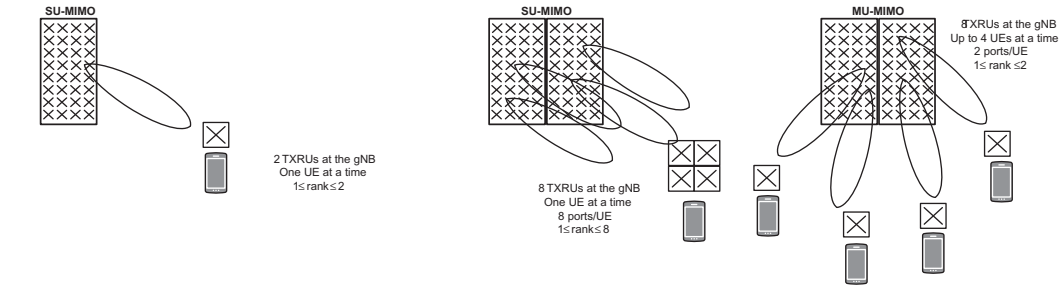


Figure 4.71
Summary of NR downlink multi-antenna operation [26].

inter-stream interference would be inevitable, and it would be the dominant factor to degrade the demodulation performance. In this case, a nonlinear receiver may be needed. When the received signal power difference among the streams is larger than 3 dB, the codeword−level successive interference cancellation or SIC receiver can satisfactorily perform. However, when the received signal power is almost the same for different data streams, the performance of SIC will deteriorate. In that case, a parallel interference cancellation receiver would be a more suitable choice. In NR, the improved DM-RSs and tracking reference signals can be used to estimate the timing offset and frequency errors in above 6 GHz frequencies [59]. The NR downlink multi-antenna operation has been discussed in different sections of this chapter and is summarized in Fig. 4.71.

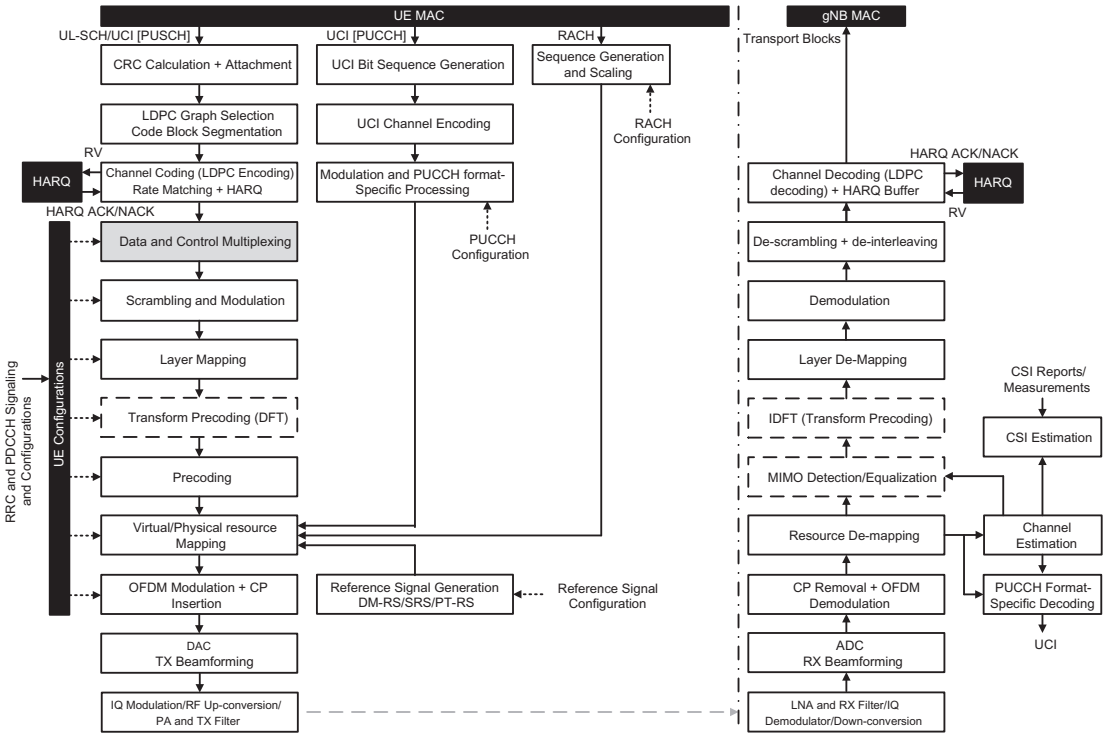


Figure 4.72
Overall uplink physical layer processing [5].

4.2 Uplink Physical Layer Functions and Procedures

4.2.1 Overall Description of Uplink Physical Layer

The NR uplink physical-layer consists of higher layer configurable functional blocks that are configured according to the uplink physical channel characteristics, use case, deployment scenario, etc. As shown in Fig. 4.72, the physical-layer processing generally includes receiving higher layer data (e.g., MAC PDUs in the case of uplink shared channel); CRC calculation and attachment; channel encoding and rate matching; modulation; mapping to physical resources and antennas; multi-antenna processing; and support of layer-1 control and HARQ-related signaling. The physical-layer model for RACH transmission is characterized by a physical RACH (PRACH) preamble format that consists of a cyclic prefix, a preamble, and a guard time during which no signal is transmitted.

4.2.2 Reference Signals

4.2.2.1 Demodulation Reference Signals

Uplink DM-RSs are used for channel estimation and, as shown in Fig. 4.73, they are subject to the same precoding as uplink shared channel. Uplink reference signals are required to have small power variation in the frequency domain to allow similar channel-estimation quality for all frequencies spanned by the reference signals. This requirement is fulfilled for CP-OFDM waveform by using a pseudo-random sequence with good autocorrelation properties. However, for DFT-precoded OFDM waveform, limited power variations as a function of time are also important to achieve signal transmission with a low cubic metric. Furthermore, sufficient number of reference signal sequences of a given length, corresponding to a certain reference signal bandwidth, should be available in order to avoid restrictions when scheduling multiple devices in different cells. It is shown that Zadoff–Chu sequences can satisfy these requirements. From a Zadoff–Chu sequence with a given group index and sequence index, additional reference signal sequences can be generated by applying different linear phase rotations in the frequency domain.

The DFT-precoded OFDM waveform in the uplink only supports single-layer transmission and is primarily specified for limited link-budget scenarios. Due to the importance of low cubic metric property and the corresponding high-power–amplifier efficiency, the reference signal structure is somewhat different compared to the CP-OFDM uplink case. In general, transmitting reference signals that are frequency-multiplexed with other uplink transmissions from the same device is not suitable for the uplink as it would negatively impact the device power-amplifier efficiency due to increased cubic metric. Instead, certain OFDM

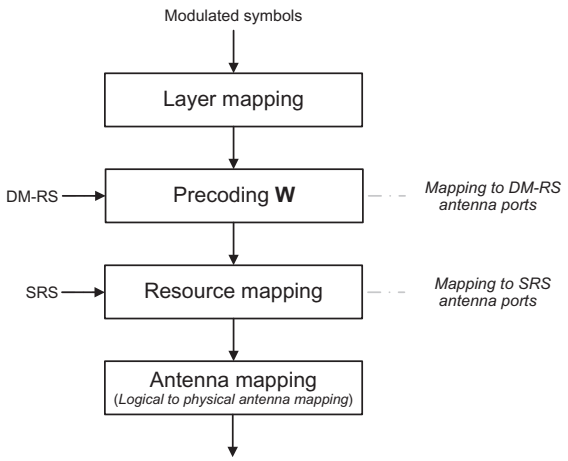


Figure 4.73
Processing of uplink DM-RS with PUSCH.

symbols within a slot are used exclusively for DM-RS transmission, that is, the reference signals are time-multiplexed with the data transmitted on PUSCH from the same device. The NR uses the same DM-RS structure for downlink and uplink in the case of CP-OFDM waveform. For DFT-spread OFDM waveform in the uplink, the DM-RS is based on Zadoff−Chu sequences and supports contiguous allocations and single-layer transmission, similar to LTE, in order to improve the power-amplifier efficiency. Multiple orthogonal reference signals can be generated in each DM-RS occasion where different reference signals are separated in the time, frequency, and code domains. Two different types of DM-RSs can be configured, namely, DM-RS Type 1 and Type 2, which differ in the maximum number of orthogonal reference signals and the mapping to the resource elements in the frequency domain. Type 1 provides up to four orthogonal reference signals using a single-symbol DM-RS and up to eight orthogonal reference signals using a double-symbol DM-RS, whereas Type 2 provides 6 and 12 patterns for single and double-symbol DM-RS, respectively. The DM-RS type 1 or 2 should not be confused with the mapping type A or B (see Section 4.1.2.1), as different mapping types can be combined with different reference signal types. In the following sections, we describe the DM-RSs corresponding to PUSCH and various PUCCH formats for the CP-OFDM uplink waveform.

4.2.2.1.1 PUSCH DM-RS

The DM-RS sequence for PUSCH $r_{DM-RS}(m)$ is generated according to $r_{DM\text{-}RS}(m) = [(1-2c(2m)) + j(1-2c(2m+1))]/\sqrt{2}$ where $c(i)$ is a length-31 Gold sequence generated by the pseudo-random sequence generator defined in [6] and initialized with $c_{init} = [2^{17}(N_{slot}^{symbol} n_{slot} + l + 1)(2N_{ID}^{n_{SCID}} + 1) + 2N_{ID}^{n_{SCID}} + n_{SCID}] \bmod 2^{31}$, l is the OFDM symbol number within the slot, n_{slot} is the slot number within a frame, and $N_{ID}^0, N_{ID}^1 \in \{0, 1, \ldots, 65535\}$ are given by the higher layer parameters *scramblingID0* and *scramblingID1* in the *DMRS-UplinkConfig* information element, respectively. The parameter $n_{SCID} \in \{0, 1\}$ may be signaled through the DM-RS initialization field of DCI format 0_1, otherwise, $n_{SCID} = 0$ [6].

The sequence $r_{DM\text{-}RS}(n)$ is mapped to an intermediate quantity $b_{k,l}^{(p,\mu)}$ according to $b_{k,l}^{(p,\mu)} = w_f(k')w_t(l')r_{DM\text{-}RS}(2n + k')$ where $k = 4n + 2k' + \Delta$ for DMR-RS Type 1 and $k = 6n + k' + \Delta$ for DMR-RS Type 2, and $k' = 0, 1$; $l = \bar{l} + l'$; $n \in \mathbb{N}$; $p = 0, 1, \ldots, N_{AP} - 1$ [6]. The spreading sequences $w_f(k')$, $w_t(l')$, and the offset Δ are given in Table 4.22 [6].

The intermediate quantity $b_{k,l}^{(p,\mu)}$ is precoded and multiplied with the amplitude scaling factor $\beta_{DM\text{-}RS}^{PUSCH}$ in order to adjust the transmit power of the reference signals and are then mapped to the physical resources according to $\left[a_{k,l}^{(p_0,\mu)} \ldots a_{k,l}^{(p_{\rho-1},\mu)}\right]^T = \beta_{DM\text{-}RS}^{PUSCH}\mathbf{W}\left[b_{k,l}^{(0,\mu)}(m) \ldots b_{k,l}^{(N_{AP}-1,\mu)}(m)\right]^T$. In the latter equation, $\mathbf{W}$ is the precoding matrix, and

Table 4.22: Parameters for PUSCH DM-RS Types 1 and 2 [6].

p	CDM Group	Δ	$w_f(k')$		$w_t(l')$	
			$k'=0$	$k'=1$	$l'=0$	$l'=1$
DM-RS Type 1						
0	0	0	+1	+1	+1	+1
1	0	0	+1	−1	+1	+1
2	1	1	+1	+1	+1	+1
3	1	1	+1	−1	+1	+1
4	0	0	+1	+1	+1	−1
5	0	0	+1	−1	+1	−1
6	1	1	+1	+1	+1	−1
7	1	1	+1	−1	+1	−1
DM-RS Type 2						
0	0	0	+1	+1	+1	+1
1	0	0	+1	−1	+1	+1
2	1	2	+1	+1	+1	+1
3	1	2	+1	−1	+1	+1
4	2	4	+1	+1	+1	+1
5	2	4	+1	−1	+1	+1
6	0	0	+1	+1	+1	−1
7	0	0	+1	−1	+1	−1
8	1	2	+1	+1	+1	−1
9	1	2	+1	−1	+1	−1
10	2	4	+1	+1	+1	−1
11	2	4	+1	−1	+1	−1

it is assumed that the resource elements $b_{k,l}^{(p,\mu)}$ are within the common resource blocks allocated for PUSCH transmission. It is further assumed that the reference point for the frequency index k is subcarrier 0 in common resource block 0; the reference point for time index l and the position l_0 of the first DM-RS symbol depends on the mapping type, that is, for PUSCH mapping type A, l is defined relative to the start of the slot, if frequency hopping is disabled; otherwise, it is measured relative to the start of each hop, and l_0 is given by the higher layer parameter *dmrs-TypeA-Position*. For PUSCH mapping type B, l is defined relative to the start of the scheduled PUSCH resources, if frequency hopping is disabled and relative to the start of each hop otherwise and $l_0 = 0$ [6].

In other words, the pseudo-random sequence corresponding to DM-RS Type 1 is mapped to every second subcarrier in the frequency domain over the OFDM symbol assigned to DM-RS transmission. As shown in Table 4.22, antenna ports 0 and 1 are associated with CDM group 0 and mapped to even-numbered subcarriers and are separated in the code-domain using different orthogonal sequences. The DM-RS corresponding to PUSCH for antenna ports 2 and 3 belong to CDM group 1 and are generated in the same way using odd-numbered subcarriers which are separated in the code domain. If more than four

orthogonal antenna ports are needed, two consecutive OFDM symbols are used. The above DM-RS structure is used over each OFDM symbol and a length-2 orthogonal sequence is applied across time, resulting in up to eight orthogonal sequences.

DM-RS Type 2 has a similar structure as Type 1; however, there are some differences, most notably the number of antenna ports supported. As shown in Fig. 4.74, each CDM group for DM-RS Type 2 consists of two neighboring subcarriers over which a length-2 orthogonal sequence is applied to separate the two antenna ports sharing the same set of subcarriers. Two such pairs of subcarriers are used in each resource block for one CDM group. Since there are 12 subcarriers in a resource block, up to three CDM groups each with two orthogonal reference signals can be created using one resource block over one

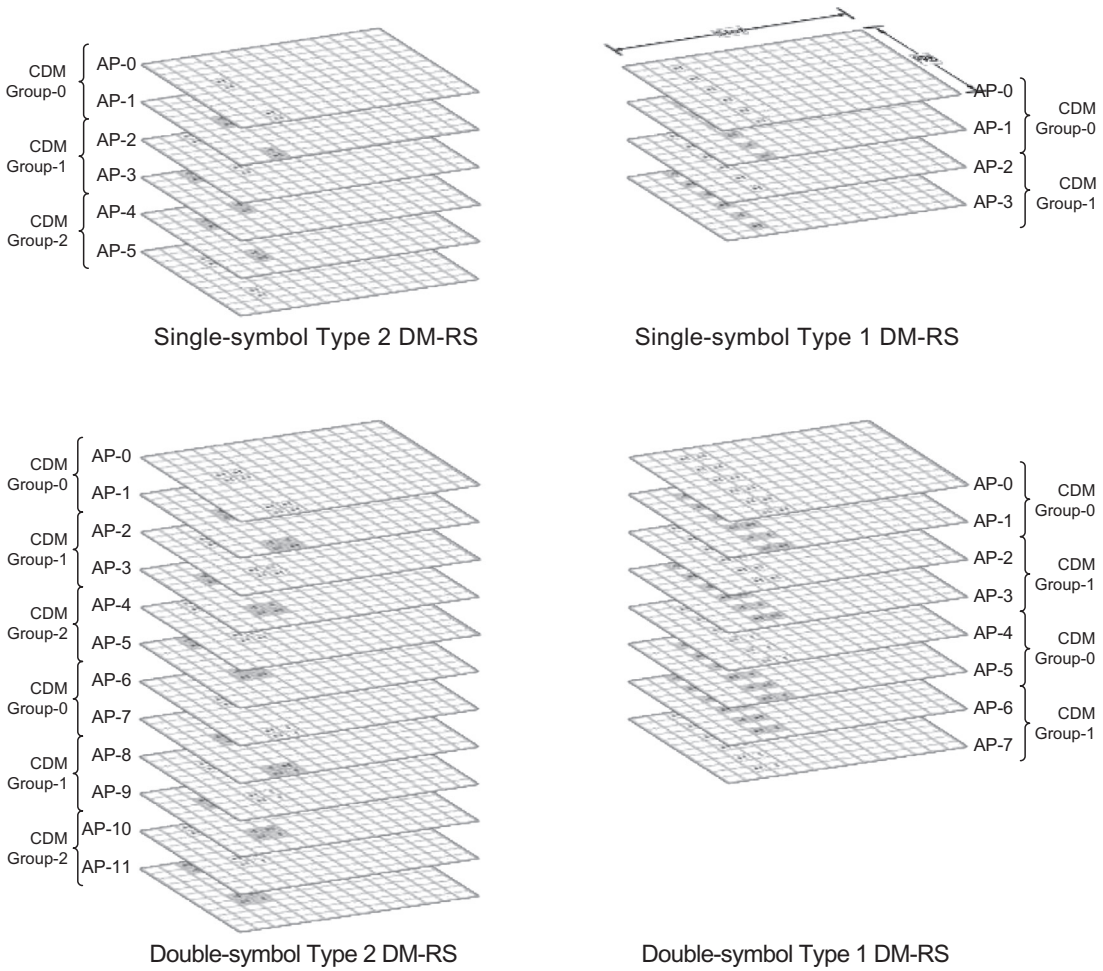


Figure 4.74
Illustration of DM-RS Type-1 and Type-2 time-frequency-code structures [14].

OFDM symbol. The number of Type 2 DM-RSs can be increased up to 12 by applying a length-2 orthogonal sequence across time domain. While the basic structures of DM-RS Type 1 and Type 2 are similar, the frequency-domain density of DM-RS Type 1 is higher than that of Type 2; however, Type 2 provides a larger number of orthogonal patterns, which useful for MU-MIMO use cases. The type of the reference signal structure is determined by dynamic scheduling and higher layer configuration.

4.2.2.1.2 PUCCH DM-RS

The DM-RS sequence for PUCCH format 1 is defined by $z(m'N_{sc}^{RB}N_{SF,0}^{PUCCH,1} + mN_{sc}^{RB} + n) = w_i(m)r_{u,v}^{(\alpha,\delta)}(n)$ where $n = 0, 1, \ldots, N_{sc}^{RB} - 1$, $m = 0, 1, \ldots, N_{SF,m'}^{PUCCH,1} - 1$, and $m' = 0$ when intra − slot frequency hopping is enabled; otherwise, $m' = 0, 1$. In the latter equation, the number of the DM-RS symbols $N_{SF,m'}$, orthogonal sequence $w_i(m)$, and the pseudo-random sequence $r_{u,v}^{(\alpha,\delta)}(n)$ are given in [6]. The DM-RS sequence is multiplied by the scaling factor $\beta_{PUCCH,1}$ in order to adjust the transmit power and is sequentially mapped starting with $z(0)$ to resource elements (k, l) in a slot on a single-antenna port such that $a_{k,l}^{(p,\mu)} = \beta_{PUCCH,1}z(m)$; $l = 0, 2, 4, \ldots$ wherein $l = 0$ corresponds to the first OFDM symbol of the PUCCH transmission and (k, l) is within the resource blocks assigned for PUCCH transmission [8].

The DM-RS sequence corresponding to PUCCH format 2, $r_{DM\text{-}RS}(m, l)$ is generated as $r_{DM\text{-}RS}(m, l) = \{[1 - 2c(2m)] + j[1 - 2c(2m + 1)]\}/\sqrt{2}$ where the pseudo-random sequence $c(i)$ is initialized with $c_{init} = [2^{17}(N_{slot}^{symbol}n_{slot} + l + 1)(2N_{ID}^0 + 1) + 2N_{ID}^0] \bmod 2^{31}$. Index l is the OFDM symbol number within the slot, n_{slot} is the slot number within the radio frame, and $N_{ID}^0 = \{0, 1, \ldots, 65535\}$. The PUCCH format 2 DM-RS sequence scaled with the scaling factor $\beta_{PUCCH,2}$ to adjust the transmit power [8] and is sequentially mapped starting with $r(0)$ to resource elements (k, l) in a slot on a single-antenna port such that $a_{k,l}^{(p,\mu)} = \beta_{PUCCH,2}r_{DM-RS}(m, l)$; $k = 3m + 1$ where the frequency index k is defined relative to subcarrier 0 of common resource block 0 and (k, l) is within the resource blocks assigned for PUCCH transmission [6].

The DM-RS sequence corresponding to PUCCH format 3/4 $r_{DM\text{-}RS}(m, l)$ is generated according to $r_{DM\text{-}RS}(m, l) = r_{u,v}^{(\alpha,\delta)}(m); m = 0, 1, \ldots, M_{sc}^{PUCCH,s} - 1$ where the number of subcarriers for this PUCCH format $M_{sc}^{PUCCH,s}$ and the pseudo-random $a_{k,l}^{(p,\mu)}$ are given in [6]. The cyclic shift α varies with the symbol number and slot number with $m_0 = 0$ for PUCCH format 3. The PUCCH format 3/4 DM-RS sequence is multiplied by the scaling factor $\beta_{PUCCH,s}$ wherein $s \in \{3, 4\}$ to adjust the transmit power and is sequentially mapped starting with $r_{DM\text{-}RS}(0)$ to resource elements (k, l) on a single-antenna port such that $a_{k,l}^{(p,\mu)} = \beta_{PUCCH,s}r_{DM\text{-}RS}(m, l)$; $m = 0, 1, \ldots, M_{sc}^{PUCCH,s} - 1$. The frequency index k is defined relative to subcarrier 0 of the lowest numbered resource block assigned for PUCCH transmission,

Table 4.23: DM-RS positions for PUCCH format 3 and 4 [6].

PUCCH Length	DM-RS Position / within PUCCH Allocation			
	No Additional DM-RS		Additional DM-RS	
	No Hopping	Hopping	No Hopping	Hopping
4	1	0,2	1	0,2
5	0,3		0,3	
6	1,4		1,4	
7	1,4		1,4	
8	1,5		1,5	
9	1,6		1,6	
10	2,7		1,3,6,8	
11	2,7		1,3,6,9	
12	2,8		1,4,7,10	
13	2,9		1,4,7,11	
14	3,10		1,5,8,12	

and time index l is given in Table 4.23 with and without intraslot frequency hopping as well as with and without additional DM-RS, where $l = 0$ corresponds to the first OFDM symbol of the PUCCH transmission [6].

4.2.2.2 Phase-Tracking Reference Signals

There is typically a mismatch between the oscillator frequencies of the transmitter and receiver in a communication system, resulting in a shift of the received signal spectrum at the baseband. In OFDM, this effect creates a misalignment between the FFT bins and the peaks of the sinc(.) pulses of the received signal, which would compromise the orthogonality between the subcarriers and would result in a spectral leakage. Each subcarrier interferes with other subcarriers, although the effect is dominant between adjacent subcarriers. Since there are many subcarriers, this is a random process is equivalent to a noise with Gaussian distribution. This random frequency offset degrades the SINR of the receiver. Therefore, an OFDM receiver will need to track and compensate the phase noise. The PT-RS is mainly introduced to compensate the CPE; however, PT-RS can also be used for ICI mitigation in higher frequency bands and potentially for CFO and Doppler estimation [54].

There is a trade-off between phase tracking accuracy and signaling overhead. If the density of PT-RS is high, phase tracking accuracy is high, and the CPE can be better compensated to achieve better performance. However, higher PT-RS density also means larger signaling overhead, which might lead to lower spectral efficiency or effective transmission rate. The studies show that the reduction of PT-RS density in time domain will degrade the BLER performance regardless of modulation order. However, the performance degradation is particularly significant when time density is reduced from 1 to 2, that is, from PT-RS for each OFDM symbol

to PT-RS for every other OFDM symbol in the case of 256 QAM [54]. In that case, although signaling overhead is halved for time density 2, that is, more information bits can be transmitted in each RB, the effective data transmission rate suffers performance loss due to degraded BLER. In contrast, for 64QAM the degradation due to time density reduction is much less. Therefore, the time density of PT-RS can be a function of modulation order and it should increase with higher modulation order.

Since the receiver only needs to track the phase difference using the PT-RS, it does not need to know the amplitude of the PT-RS. Therefore, unlike other reference signals, for example, DM-RS, where the reference signals are formed by a pseudo-random sequence of symbols, PT-RS reference signals can use the exact same symbol. In MU-MIMO, PT-RS can be configured for each user and it is possible that the same subcarrier is used for multiple users, thus PT-RS collisions may happen, which would degrade the CPE compensation for two reasons: (1) the interference pattern is not completely random since the same symbol is used for PT-RS and (2) the interference level is higher when the power of PT-RS is boosted for more accurate CPE compensation. In such cases, it would be better to avoid PT-RS collision so that the interference is randomized without power boosting. This can be avoided by introducing an RB-level offset when configuring PT-RS for each user [54].

In NR, both DFT-S-OFDM and CP-OFDM waveforms are supported for the uplink transmissions and PT-RS signals are necessary for both waveforms. The PT-RS insertion follows a common framework for both downlink and uplink in case of CP-OFDM waveforms. In the case of DFT-S-OFDM, two types of insertion mechanisms for PT-RSs were studied, that is, pre-DFT and post-DFT insertion. In the former mechanism, PT-RS signals are inserted in the frequency domain before DFT precoding so that the resulting waveform still maintains the single-carrier properties. In the latter mechanism, the PT-RS are inserted after DFT precoding of the data symbols via various mechanisms such as puncturing. The PAPR of such a mechanism can however be controlled by using some signal processing techniques. In the time domain, the PT-RS locations can be configured to be either present in every symbol or every other symbol. The NR supports pre-DFT PT-RS insertion to conserve the single carrier property.

In the frequency domain, PT-RSs are transmitted in every second or fourth resource block, resulting in a sparse frequency-domain structure. The density in the frequency domain corresponds to the scheduled transmission bandwidth such that the higher the bandwidth, the lower the PT-RS density. For the smallest bandwidths, no PT-RS is transmitted. To reduce the risk of collisions between PT-RSs associated with different devices scheduled on overlapping time-frequency resources, the subcarrier number and the resource blocks used for PT-RS transmission are determined by the C-RNTI of the UE. The antenna port used for PT-RS transmission is given by the lowest numbered antenna port in the DM-RS antenna port group [14].

In a multi-TRP deployment, a UE can be supported by multiple co-located or non-co-located transmission points belonging to the same or different gNBs. The NR further supports large number of antenna elements at the gNB. These antenna elements are typically grouped as

panels, where the signals feeding the antenna panels are generated by separate oscillators, which require individual phase noise compensation. In designing PT-RS for multi-TRP deployments, the orthogonal time-frequency allocation of PT-RS is crucial to minimize interference. The orthogonality of the PT-RS can be ensured in the frequency domain via frequency-division multiplexing of the PT-RS bearing resource elements. The increase in signaling overhead is a valid concern when there is a higher number of gNBs or many transmit panels. The typical CPE caused by the phase noise rotates the constellations by a limited margin, so only the higher order modulation schemes are impacted by the CPE. The users with higher MCS receive good SNR levels and are usually located closer to the gNB. When the MU-MIMO users are grouped, there will be higher and lower MCS users in these groups. If the PT-RS is transmitted without power boosting and with a wider beam than the narrow-beam data transmissions, the received EIRP for the PT-RS will be lower relative to the data transmissions. The lower MCS users that are generally further away from the cell center will receive PT-RS with a much lower effective power and will be able to discard PT-RS as interference, that is, they will not need CPE correction. They will be able to request the gNB to allocate data within these resource elements transmitted through narrow-beams. The same resource elements are used for the PT-RS in the wider beam transmissions for the benefit of higher MCS users, for whom the same resource elements will not be utilized in the narrow-beam data transmissions. With this effective power discrimination, non-orthogonal multiplexing of PT-RS and data is possible for the MU-MIMO configurations, which effectively increases the system spectral efficiency [54].

In CP-OFDM uplink, the precoded phase-tracking reference signal for subcarrier k on layer j is given by $r_{PT\text{-}RS}(p,m) = [(1-2c(2m)) + j(1-2c(2m+1))]/\sqrt{2}$ if $p = p'$ or $p = p''$ where antenna ports p' or (p', p'') are associated with PT-RS transmission. The pseudo-random sequence $c(i)$ is initialized with $c_{init} = [2^{17}(N_{slot}^{symbol} n_{slot} + l + 1)(2N_{ID}^{n_{SCID}} + 1) + 2N_{ID}^{n_{SCID}} + n_{SCID}] \bmod 2^{31}$, in which l is the OFDM symbol number within the slot, n_{slot} is the slot number within a frame, and $N_{ID}^0, N_{ID}^1 \in \{0, 1, \ldots, 65535\}$ are given by the higher layer parameters *scramblingID0* and *scramblingID1*, respectively. The parameter $n_{SCID} \in \{0, 1\}$ may be signaled through the DM-RS initialization field of DCI format 0_1; otherwise, $n_{SCID} = 0$ [6].

The UE transmits PT-RSs (if configured) only in the resource blocks designated to PUSCH transmission. The PT-RS is mapped to resource elements according to $[a_{k,l}^{(p_0,\mu)} \cdots a_{k,l}^{(p_{\rho-1},\mu)}]^T = \beta_{PT\text{-}RS}\mathbf{W}[r_{PT\text{-}RS}(p_0, 2n+k') \cdots r_{PT\text{-}RS}(p_{v-1}, 2n+k')]^T$ where $k = 4n + 2k' + \Delta$ for configuration Type 1 or $k = 6n + k' + \Delta$ for configuration Type 2, if l is within the OFDM symbols allocated for the PUSCH transmission, the resource element (k, l) is not used for DM-RS. The parameters k' and Δ as well as the precoding matrix $\mathbf{W}$ are given in [6]. The configuration type is provided by the higher layer parameter *DMRS-UplinkConfig*. In the preceding expression, scaling factor $\beta_{PT\text{-}RS}$ is used to adjust the transmit power. The set of time indices l is defined relative to the start of the PUSCH allocation is defined by $\max(l_{ref} + (i-1)L_{PT\text{-}RS} + 1, l_{ref}), \ldots, l_{ref} + iL_{PT\text{-}RS} \forall L_{PT\text{-}RS} \in \{1, 2, 4\}$ where any symbol in this interval which overlaps with a DM-RS symbol is skipped.

Table 4.24: Time-domain/frequency-domain density of phase tracking reference signal as a function of scheduled modulation coding scheme/ bandwidth [9].

Scheduled MCS	Time-Domain Density $L_{PT\text{-}RS}$
$0 \leq \text{MCS} < \text{MCS}_1$	No PT-RS
$\text{MCS}_1 \leq \text{MCS} < \text{MCS}_2$	Every OFDM symbol
$\text{MCS}_2 \leq \text{MCS} < \text{MCS}_3$	Every second OFDM symbol
$\text{MCS}_3 \leq \text{MCS} < \text{MCS}_4$	Every fourth OFDM symbol
Scheduled Bandwidth	**Frequency Domain Density $K_{PT\text{-}RS}$**
$0 \leq N_{RB} < N_{RB_1}$	No PT-RS
$N_{RB_1} \leq N_{RB} < N_{RB_2}$	Every second RB
$N_{RB_2} \leq N_{RB}$	Every fourth RB

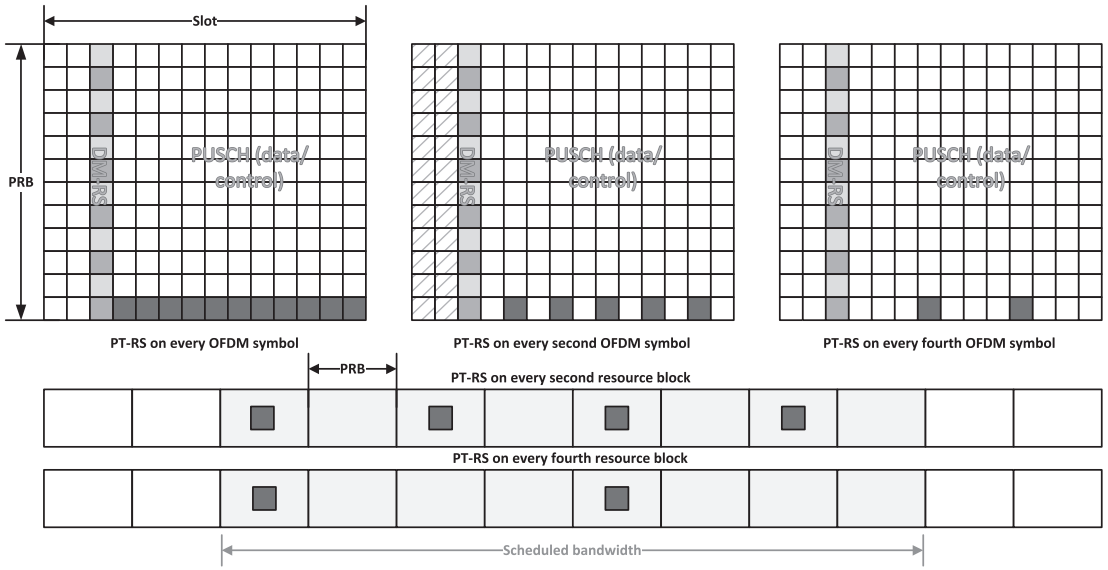


Figure 4.75
Example CP-OFDM uplink time-frequency resource mappings.

The resource blocks allocated for PUSCH transmission are numbered from 0 to N_{RB-1} relative to the lowest scheduled resource block for the purpose of PT-RS transmission. The subcarriers associated with these resource blocks are numbered in increasing order from 0 to $N_{sc}^{RB}N_{RB} - 1$ starting with the lowest frequency. The PT-RS is mapped to subcarriers $k = k_{ref}^{RE} + (iK_{PT\text{-}RS} + k_{ref}^{RB})N_{sc}^{RB}$ where $K_{PT\text{-}RS} \in \{2, 4\}$ (see Table 4.24). The parameter $k_{ref}^{RB} = n_{RNTI} \bmod K_{PT\text{-}RS}$, if $N_{RB} \bmod K_{PT\text{-}RS} = 0$; otherwise, $k_{ref}^{RB} = n_{RNTI} \bmod (N_{RB} \bmod K_{PT\text{-}RS})$. The parameter n_{RNTI} is the RNTI associated with the DCI that is used to schedule PUSCH [6]. Example time-frequency mapping of CP-OFDM uplink PT-RS is depicted in Fig. 4.75.

4.2.2.3 Sounding Reference Signal

A UE can be configured to transmit SRS in order to enable the gNB to estimate the uplink channel. Similar to the downlink CSI-RS, the SRS can serve as QCL reference for other physical channels such that they can be configured and transmitted quasi-co-located with SRS. As a result, if the knowledge of a suitable receive beam for the SRS is available, the receiver would know that the same receive beam should be suitable for other physical channels. The SRS supports up to four antenna ports, and it is designed to have low cubic metric, enabling efficient operation of the high-power amplifier. In general, the SRS can span one, two, or four consecutive OFDM symbols and is located within the last six symbols of a slot. In the frequency domain, an SRS occasion has a comb structure where the SRS is transmitted on every Nth subcarrier where $N = 2$ or 4, referred to as comb-2 or comb-4. The SRS transmissions from different devices can be frequency-multiplexed within the same frequency range using different comb patterns corresponding to different frequency offsets. For comb-2, that is, transmitting SRS on every other subcarrier, two SRSs can be frequency multiplexed, whereas for comb-4, up to four SRSs can be frequency multiplexed. Fig. 4.76 illustrates example SRS multiplexing assuming a comb-2 structure spanning two OFDM symbols.

The sequences used to represent a set of SRS are based on Zadoff–Chu sequences.[19]

[19] A Zadoff–Chu sequence is a complex-valued sequence with constant amplitude property whose cyclically shifted versions exhibit low cross-correlations. Thus under certain conditions, the cyclically shifted versions of each sequence remain orthogonal to one another. A Zadoff–Chu sequence that has not been shifted is referred to as a root sequence. The uth root Zadoff–Chu sequence of prime length N is defined as follows:

$$xu(n) \triangleq \begin{cases} e^{-j[\pi un(n+1)]/N} & 0 \le n < N-1 \quad (N \text{ is an odd integer}) \\ e^{-j\pi un^2/N} & 0 \le n < N-1 \quad (N \text{ is an even integer}) \end{cases}$$

where N is an integer, denoting the length of the Zadoff–Chu sequence. One can verify that $x_u(n)$ is periodic with period N, that is, $x_u(n) = x_u(n+N), \forall n$. In other words, the sequence index $u$is a prime relative to N. For a fixed value of u the Zadoff–Chu sequence has an ideal periodic autocorrelation property (i.e., the periodic autocorrelation is zero for all time shifts other than zero). For different values of index u the Zadoff–Chu sequences are not orthogonal, rather exhibit low cross-correlation. If the sequence length N is selected as a prime number, there are different sequences with periodic cross-correlation of $1/\sqrt{N}$ between any two sequences regardless of time shift. The Zadoff–Chu sequences are a subset of constant amplitude zero autocorrelation sequences. The properties of Zadoff–Chu sequences can be summarized as follows:

- They are periodic with period N, if N is a prime number, that is, $x_u(n+N) = x_u(n)$.
- Given N is a prime number, the DFT of a Zadoff–Chu sequence is another Zadoff–Chu sequence conjugated and time-scaled multiplied by a constant factor, that is, $X_u[k] = x_u^*(vk)X_u[0]$ where v is the multiplicative inverse of u modulo N. It can be shown that $x_u^*(vk) = x_v^*(k)e^{j\pi(1-v)k/N}$.
- The autocorrelation of a prime-length Zadoff–Chu sequence with a cyclically shifted version of itself also yields zero autocorrelation sequence, that is, it is non-zero only at one instant which corresponds to the cyclic shift zero.
- The cross-correlation between two prime-length Zadoff–Chu sequences, that is, different u, is constant and equal to $1/\sqrt{N}$.
- The Zadoff–Chu sequences have low PAPR.

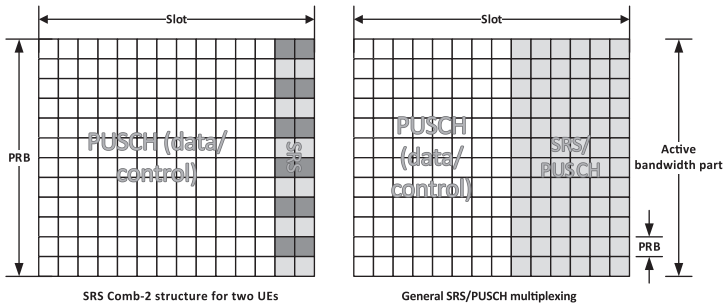


Figure 4.76
Example illustration of NR SRS structure in time and frequency domains [6].

Although Zadoff−Chu sequences of prime length are preferred in order to maximize the number of available sequences, the SRS sequences are not of prime length. The SRS sequences are extended Zadoff−Chu sequences based on the longest prime-length Zadoff−Chu sequence with a length N less than or equal to the desired SRS sequence length. The sequence is then cyclically extended in the frequency domain up to the desired SRS sequence length. As the extension is performed in the frequency domain, the extended sequence has a constant spectrum and a perfect cyclic autocorrelation, but the time-domain amplitude is not constant and slightly varies. The extended Zadoff−Chu sequences are used as SRS sequences for sequence lengths of 36 or larger, corresponding to an SRS extending over 6 and 12 resource blocks in the cases of comb-2 and comb-4, respectively. For shorter sequence lengths, special flat-spectrum sequences with good time-domain envelope properties are found through computer search since there would not be sufficient number of Zadoff−Chu sequences available.

An SRS resource is configured by the *SRS-Resource* information element and consists of 1, 2, or 4 antenna ports, where the number of antenna ports is set by the higher layer parameter *nrofSRS-Ports*, and $N_{symb}^{SRS} \in \{1, 2, 4\}$ consecutive OFDM symbols provided via parameter *nrofSymbols* contained in the higher layer parameter *resourceMapping*. The starting position l_0 in the time domain given by $l_0 = N_{slot}^{symb} - 1 - l_{offset}$ where the offset parameter $l_{offset} \geq N_{symb}^{SRS} - 1, l_{offset} \in \{0, 1, \ldots, 5\}$ counts symbols backwards from the end of the slot and is given by the field *startPosition* contained in the higher layer parameter *resourceMapping*. The SRS sequence for an SRS resource is generated as $r^{(p_i)}(n, l') = r_{u,v}^{(\alpha_i, \delta)}(n) \forall 0 \leq n \leq M_{sc,b}^{RS} - 1, l' \in \{0, 1, \ldots, N_{symb}^{SRS} - 1\}$ where the length of the SRS sequence is given by $M_{sc,b}^{RS} = m_{SRS,b} N_{sc}^{RB} / K_{TC}$, $r_{u,v}^{(\alpha_i, \delta)}(n)$ is a Zadoff−Chu sequence, $\delta = \log_2(K_{TC})$ in which K_{TC} denotes the transmission comb number given by the higher layer parameter *transmissionComb*. The cyclic shift α_i for antenna port p_i is given

as $\alpha_i = 2\pi n_{SRS}^{cs,i}/n_{SRS}^{cs,\max}$ where $n_{SRS}^{cs,i} = (n_{SRS}^{cs} + n_{SRS}^{cs,\max}(p_i - 1000)/N_{ap}^{SRS}) \bmod n_{SRS}^{cs,\max}$ and $n_{SRS}^{cs} \in \{0, 1, \ldots, n_{SRS}^{cs,\max} - 1\}$ is provided by the higher layer parameter *transmissionComb*. The maximum number of cyclic shifts is $n_{SRS}^{cs,\max} = 12$ when $K_{TC} = 4$ and $n_{SRS}^{cs,\max} = 8$ when $K_{TC} = 2$ [6]. The sequence group is defined as $u = [f_{gh}(n_{slot}, l') + n_{ID}^{SRS}] \bmod 30$ and the sequence number v depends on the higher layer parameter *groupOrSequenceHopping* in the *SRS-Config* information element. Furthermore, the SRS sequence identity n_{ID}^{SRS} is given by the higher layer parameter *sequenceId* in the *SRS-Config* information element. If *groupOrSequenceHopping* parameter indicates neither group nor sequence hopping, $f_{gh}(n_{slot}, l') = 0$ and $v = 0$ are used; otherwise, if parameter *groupOrSequenceHopping* indicates *groupHopping*, group hopping is used and $f_{gh}(n_{slot}, l') = \left(\sum_{m=0}^{7} c\left[8(n_{slot}N_{slot}^{symb} + l_0 + l') + m\right]2^m\right) \bmod 30$ and $v = 0$ where the pseudo-random sequence $c(i)$ is initialized with $c_{init} = n_{ID}^{SRS}$ at the beginning of each radio frame. If parameter *groupOrSequenceHopping* indicates *sequenceHopping*, sequence hopping is used and $f_{gh}(n_{slot}, l') = 0$ and $v = c(n_{slot}N_{slot}^{symb} + l_0 + l')$ for $M_{sc,b}^{SRS} \geq 6N_{sc}^{RB}$; otherwise, $v = 0$. The pseudo-random sequence $c(i)$ is initialized similar to the group hopping case [6].

Each SRS is transmitted on a designated SRS resource, and the SRS sequence $r^{(p_i)}(n, l')$ for each OFDM symbol l' and antenna port p_i is multiplied by a scaling factor β_{SRS} to adjust its transmit power. The scaled sequence is then sequentially mapped to resource elements (k, l) starting with $r^{(p_i)}(0, l')$ in a slot for each antenna port such that $a_{K_{TC}k' + k_0^{(p_i)}, l' + l_0}^{(p_i)} = \beta_{SRS} r^{(p_i)}(k', l')/\sqrt{N_{ap}} \forall k' = 0, 1, \ldots, M_{sc,b}^{RS} - 1$ and $l' = 0, 1, \ldots, N_{symb}^{SRS} - 1$. The length of the SRS sequence is given by $M_{sc,b}^{RS} = m_{SRS,b}N_{sc}^{RB}/K_{TC}$ where $m_{SRS,b}$ is given in [6]. The frequency-domain starting position $k_0^{(p_i)}$ is defined by $k_0^{(p_i)} = \overline{k}_0^{(p_i)} + \sum_{b=0}^{B_{SRS}} K_{TC}M_{sc,b}^{SRS}n_b$ where B_{SRS} denotes the SRS bandwidth, $\overline{k}_0^{(p_i)} = n_{shift}N_{sc}^{RB} + k_{TC}^{(p_i)}$, $k_{TC}^{(p_i)} = (\overline{k}_{TC} + K_{TC}/2) \bmod K_{TC}$ if $n_{SRS}^{cs} \in \{n_{SRS}^{cs,\max}/2, \ldots, n_{SRS}^{cs,\max} - 1\}$ and $N_{ap}^{SRS} = 4$; otherwise, $k_{TC}^{(p_i)} = \overline{k}_{TC}$ [6]. The frequency domain shift value n_{shift} adjusts the SRS allocation with respect to the common resource block grid. The transmission comb offset $\overline{k}_{TC} \in \{0, 1, K_{TC} - 1\}$ is contained in the higher layer parameter *transmissionComb* in the *SRS-Config* information element and n_b is a frequency position index. Frequency hopping of the SRS is configured by the parameter $b_{hop} \in \{0, 1, 2, 3\}$. If $b_{hop} \geq B_{SRS}$, frequency hopping is disabled and the frequency position index n_b is set to a constant as follows $n_b = \lfloor 4n_{RRC}/m_{SRS,b} \rfloor \bmod N_b$ for all N_{symb}^{SRS} OFDM symbols of the SRS resource. The value of the parameter n_{RRC} is given by the higher layer parameter *freqDomainPosition* and the values of $m_{SRS,b}$ and N_b for $b = B_{SRS}$ are given in [6]. If $b_{hop} < B_{SRS}$, the frequency hopping is enabled and the frequency position indices n_b are defined as $n_b = \lfloor 4n_{RRC}/m_{SRS,b} \rfloor \bmod N_b$ if $b \leq b_{hop}$; otherwise, $n_b = \{F_b(n_{SRS}) + \lfloor 4n_{RRC}/m_{SRS,b} \rfloor\} \bmod N_b$ where N_b and $F_b(n_{SRS})$ are defined in [6]. The quantity n_{SRS} counts the

number of SRS transmissions. For the case of an SRS resource configured as aperiodic by the higher layer parameter *resourceType*, $n_{SRS} = \lfloor l'/R \rfloor$ within the slot where N_{symb}^{SRS} symbol SRS resource is transmitted. The quantity $R \leq N_{symb}^{SRS}$ is a repetition factor configured via higher layer signaling.

For the case of an SRS resource configured as periodic or semi-persistent by the higher layer parameter *resourceType*, the value of the SRS counter is given by $n_{SRS} = \left(N_{frame}^{slot} n_{frame} + \quad n_{slot} - T_{offset}\right)\left(N_{symb}^{SRS}/RT_{SRS}\right) + \lfloor l'/R \rfloor$ for slots that satisfy $\left(N_{frame}^{slot} n_{frame} + n_{slot} - T_{offset}\right) \bmod \ T_{SRS} = 0$, where T_{SRS} is SRS periodicity defined in the number of slots and T_{offset} is the slot offset [6]. Note that when supporting more than one SRS antenna port, different antenna ports share the same set of resource elements and the same baseline SRS sequence; nevertheless, different phase rotations are applied to separate them. Applying a phase rotation in the frequency domain is equivalent to cyclic shift in the time domain.

As we mentioned earlier, the SRS can be configured as periodic, semi-persistent, or aperiodic transmission A periodic SRS is transmitted with a certain configured periodicity and a certain configured slot offset within that period. A semi-persistent SRS has a configured periodicity and slot offset in the same way as a periodic SRS; however, the SRS transmission is performed according to the configured periodicity and slot offset that is activated or deactivated via MAC control element signaling. An aperiodic SRS is only transmitted when explicitly triggered by means of a DCI. It should be noted that SRS activation/deactivation or triggering for semi-persistent and aperiodic cases is not done for a specific SRS, rather for an SRS resource set which may include multiple SRSs [14].

A UE can be configured with one or several SRS resource sets, where each resource set includes one or more configured SRSs. All SRS occasions included within a configured SRS resource set are of the same type. In other words, periodic, semi-persistent, or aperiodic transmission is a property of an SRS resource set. A UE can be configured with multiple SRS resource sets that can be used for different purposes, including both downlink and uplink multi-antenna precoding and/or downlink and uplink beam management. The transmission of the set of configured SRS included in an aperiodic SRS resource set is triggered by a DCI. More specifically, DCI format 0_1 containing uplink grant and DCI format 1_1 containing downlink scheduling assignment include a 2-bit SRS-request that can trigger the transmission of one of the three different aperiodic SRS resource sets configured for the UE and the fourth bit combination corresponds to no trigger.

The SRS antenna ports are typically not mapped directly to the UE's physical antennas, rather via some antenna mapping scheme. In order to provide connectivity regardless of the rotational direction of the device, the NR devices supporting high-frequency operation will

typically include multiple antenna panels pointing in different directions. The SRS may be mapped to one of those panels and transmission from different panels will then correspond to different antenna mapping schemes. The antenna mapping scheme has a real impact despite the fact that it is transparent to the gNB receiver; thus it is seen as an integral part of the overall channel from the UE to the gNB. The gNB may estimate the channel based on SRS transmission from a UE and subsequently select a precoding matrix that the device should use for uplink transmission. The device is then assumed to use that precoding matrix in combination with the antenna mapping scheme that is applied to the SRS. In other cases, a device may be explicitly scheduled for data transmission using the antenna ports defined for SRS transmission. In practice, this implies that the device will transmit the data using the same antenna mapping scheme that was used for SRS transmission, meaning that the UE should use the same beam or panel that was used for SRS transmission [14].

4.2.3 Control Channels

4.2.3.1 Physical Uplink Control Channel

Physical uplink control channel carries the uplink control information (UCI) from the UEs to the gNB. The new radio has specified five PUCCH formats (as shown in Table 4.25) that are identified depending on the duration of PUCCH and the UCI payload size. The NR further supports simultaneous transmission of data and control information on PUSCH. Thus if the device is transmitting on PUSCH, the UCI is multiplexed with data on the allocated resources instead of being transmitted on PUCCH. It must be noted that simultaneous transmission of PUSCH and PUCCH is not supported in NR Rel-15. The NR PUCCH can be beamformed by configuring one or more spatial correspondence between PUCCH and downlink physical

Table 4.25: PUCCH formats [6].

Type	PUCCH Format	Number of OFDM Symbols N_{symb}^{PUCCH}	Number of Bits	Description
Short PUCCH	0	1–2	≤ 2	Short PUCCH of one or two symbols with small UCI payloads of up to 2 bits with UE multiplexing in the same PRB
	2	1–2	> 2	Short PUCCH of one or two symbols with large UCI payloads of more than 2 bits with no multiplexing in the same PRB
Long PUCCH	1	4–14	≤ 2	Long PUCCH of 4–14 symbols with small UCI payloads of up to 2 bits with multiplexing in the same PRB
	3	4–14	> 2	Long PUCCH of 4–14 symbols with medium UCI payloads with some multiplexing capacity in the same PRB
	4	4–14	> 2	Long PUCCH of 4–14 symbols with large UCI payloads with no multiplexing capacity over the same PRB

signals such as CSI-RS or SS block. As a result, the device can transmit PUCCH using the same beam as it used for receiving the corresponding downlink signal. For example, if the spatial relation between PUCCH and SS block is configured, the device will transmit PUCCH using the same beam as it used for receiving the SS block. Multiple spatial relations can be configured and selected via MAC control elements [71].

It will be shown later in this section that the short PUCCH format of up to two UCI bits is based on sequence selection, while the short PUCCH format of more than two UCI bits frequency multiplexes UCI and DM-RS. The long PUCCH formats time-multiplex the UCI and DM-RS. Frequency hopping is supported for long PUCCH formats and for short PUCCH formats of duration two symbols. Long PUCCH formats can be repeated over multiple slots. The UCI multiplexing in PUSCH is further supported when UCI and PUSCH transmissions coincide in the same slot, that is, UCI carrying HARQ-ACK feedback with 1 or 2 bits is multiplexed by puncturing and rate-matching PUSCH. In all other cases, the UCI is multiplexed by rate matching PUSCH. The UCI may carry CSI, HARQ ACK/NACK, or scheduling request (SR). The QPSK modulation is used for long PUCCH with 2 or more bits of information, and short PUCCH with more than 2 bits of information. The BPSK modulation is used for long PUCCH with a single information bit. Transform precoding is applied to long PUCCH [11].

In deployment scenarios where a SUL is used the UE is configured with two uplink carriers and one downlink carrier in the same cell. The transmissions on those uplink carriers are controlled by the network to avoid overlapping PUSCH/PUCCH transmissions in the time domain. The overlapping transmissions on PUSCH are avoided with properly scheduling the uplink transmissions, while overlapping transmissions on PUCCH are circumvented by proper configuration, that is, PUCCH can only be configured for one of the two uplink carriers of the cell. In addition, initial access is supported on each of those uplink carriers [11].

A UE is semi-statically configured via RRC signaling to perform periodic CSI reporting using PUCCH and can be configured for multiple periodic CSI reports corresponding to one or more CSI reporting setting indications, where the associated CSI measurement links and CSI Resource Settings are also configurable. Periodic CSI reporting on PUCCH formats 2, 3, 4 supports Type I CSI with wideband granularity. A UE performs semi-persistent CSI reporting on PUCCH after successfully decoding a selection command, which contains one or more reporting setting indications where the associated CSI measurement links and CSI resource settings are configured. Semi-persistent CSI reporting on PUCCH supports Type I CSI. The semi-persistent CSI reporting on PUCCH format 2 supports Type I CSI with wideband granularity, whereas semi-persistent CSI reporting on PUCCH format 3 or 4 supports Type I subband CSI and Type II CSI with wideband frequency granularity. When PUCCH carry Type I CSI with wideband frequency granularity, the CSI payloads carried by PUCCH format 2, and PUCCH format 3 or 4 are identical irrespective of RI and CRI. For

Type I CSI subband reporting on PUCCH format 3 or 4, the payload is split into two parts. The first part may contain RI, CRI, and/or CQI for the first codeword. The second part contains PMI and the CQI for the second codeword when $RI > 4$. A semi-persistent report carried on PUCCH format 3 or 4 supports only part 1 of Type II CSI feedback. Supporting Type II CSI reporting on PUCCH format 3 or 4 is considered a UE capability. A Type II CSI report (part 1 only) is carried on PUCCH format 3 or 4 and is calculated independent of any Type II CSI reports carried on PUSCH. When the UE is configured with CSI reporting on PUCCH format 2, 3, or 4, each PUCCH resource is configured for each candidate uplink BWP. The UE will never report CSI with a payload size larger than 115 bits when configured with PUCCH format 4 [8].

The NR physical uplink control channel is used for transmission of HARQ ACK/NACK for received downlink data; CSI feedback related to the downlink channel conditions to assist dynamic scheduling, and SR indicates that a UE needs uplink resources for data transmission. The NR PUCCH supports two transmission modes (see Fig. 4.77):

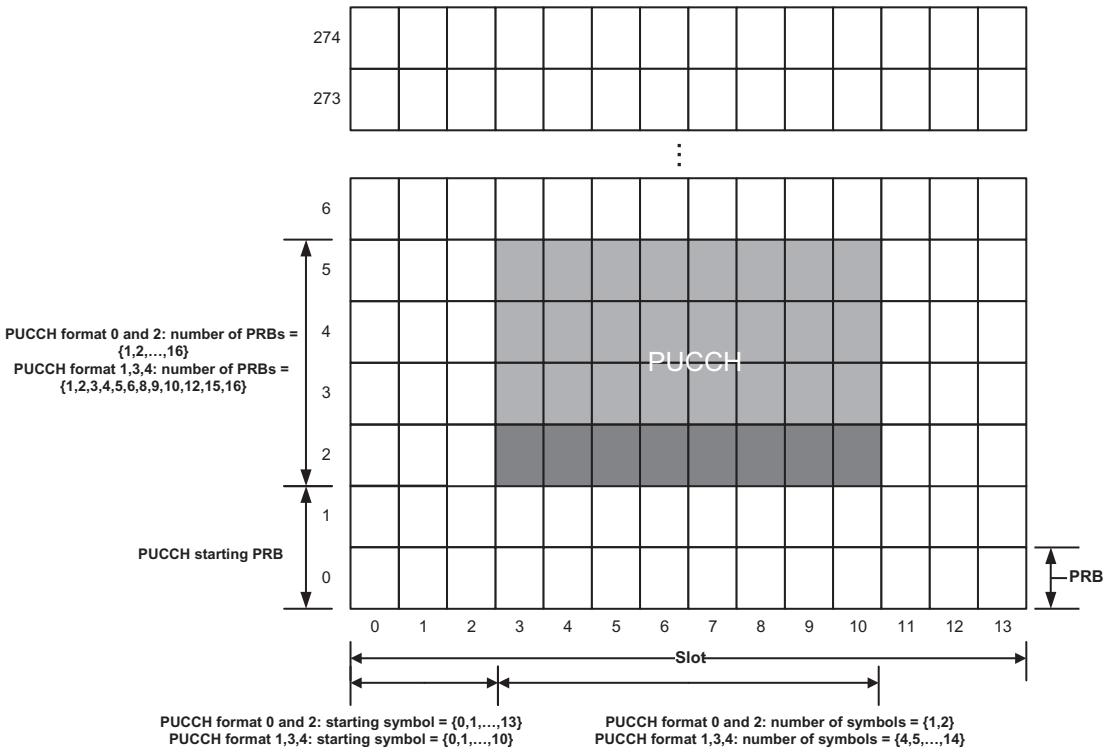


Figure 4.77
General time-frequency structure of PUCCH for various formats [6].

- *Short-PUCCH*: NR PUCCH can be transmitted in short duration around the last uplink symbol(s) of a slot. The transmission can span one or two OFDM symbols.
- *Long-PUCCH*: NR PUCCH can be transmitted in long duration at least over four uplink symbols to improve coverage. It is also considered to support a long-PUCCH transmission over multiple slots.

One or multiple PRBs can be allocated as the minimum resource unit size in frequency domain for short-PUCCH and long-PUCCH. Intra-slot frequency hopping can be configured for PUCCH format 1, 3, or 4, that is, the long-PUCCH, where the number of symbols in the first hop is given by $\left\lfloor N_{symb}^{PUCCH}/2 \right\rfloor$ in which N_{symb}^{PUCCH} is the length of PUCCH transmission in the number of OFDM symbols. PUCCH formats 0, 1, 3, and 4 use the low-PAPR Zadoff−Chu sequences $r_{u,v}^{(\alpha,\delta)}(n)|_{\delta=0}$ where α is a cyclic shift of a base sequence $\overline{r}_{u,v}(n)$ such that $r_{u,v}^{(\alpha,0)}(n) = e^{j\alpha n}\overline{r}_{u,v}(n)$, $0 \le n < M_{ZC}$ in which $M_{ZC} = 12m$ is the length of the sequence (assuming $\delta = 0$). Multiple sequences can be generated from a single base sequence based on different values of α (cyclic shift). Base sequences $\overline{r}_{u,v}(n)$ are divided into groups, where $u \in \{0, 1, \ldots, 29\}$ is the group number and v is the base sequence number within the group, such that each group contains one base sequence $v = 0$ each of length $M_{ZC} = 12m$, $1 \le m \le 5$ and two base sequences $v = 0, 1$ each of length $M_{ZC} = 12m$, $m \ge 6$. The definition of the base sequence $\overline{r}_{u,v}(0), \ldots, \overline{r}_{u,v}(M_{ZC} - 1)$ depends on the sequence length M_{ZC} [6]. The sequence group $u = (f_{gh} + f_{ss}) \bmod 30$ and the sequence number v within the group depend on the value of the RRC parameter *pucch-GroupHopping*. For example, if RRC parameter *pucch-GroupHopping* is set to "enabled," $f_{gh} = \left(\sum_{m=0}^{7} 2^m c[8(2n_{slot} + n_{hop}) + m]\right) \bmod 30$; $f_{ss} = n_{ID} \bmod 30$; $v = 0$ where n_{ID} is the hopping identifier identified by RRC signaling, and $c(n)$ is a pseudo-random sequence which was defined earlier. The frequency hopping index $n_{hop} = 0$, if frequency hopping is disabled; otherwise, $n_{hop} = 0$ for the first hop and $n_{hop} = 1$ for the second hop. The cyclic shift α is a function of the symbol and slot number and is expressed as $\alpha_l = ([m_0 + m_{cs} + n_{cs}(n_{slot}, l + l')] \bmod 12)\pi/6$ where the parameters are defined in [6]. In the latter equation the function $n_{cs}(n_{slot}, l)$ is defined as $n_{cs}(n_{slot}, l) = \sum_{m=0}^{7} 2^m c(8N_{slot}^{symb} n_{slot} + 8l + m)$ in which the pseudo-random sequence $c(i)$ was defined earlier and the pseudo-random sequence generator is initialized with the RRC parameter *hoppingId*. The general time-frequency structure of PUCCH for various formats is depicted in Fig. 4.77.

4.2.3.1.1 PUCCH Format 0 Structure and Physical Processing

PUCCH format 0 is a short PUCCH format that can transport up to 2 bits. It is used for HARQ-ACK and SRs. As shown in Fig. 4.78, PUCCH format 0 sequence is generated $x(12l + n) = r_{u,v}^{(\alpha,\delta)}(n); n = 0, 1, \ldots, 11$ and $l = 0$ for single-symbol PUCCH transmission and $l = 0, 1$ for double-symbol PUCCH transmission. The sequence $r_{u,v}^{(\alpha,\delta)}(n)$ was defined in the previous section, in which the parameter m_{cs} depends on the UCI. The sequence $x(n)$ is

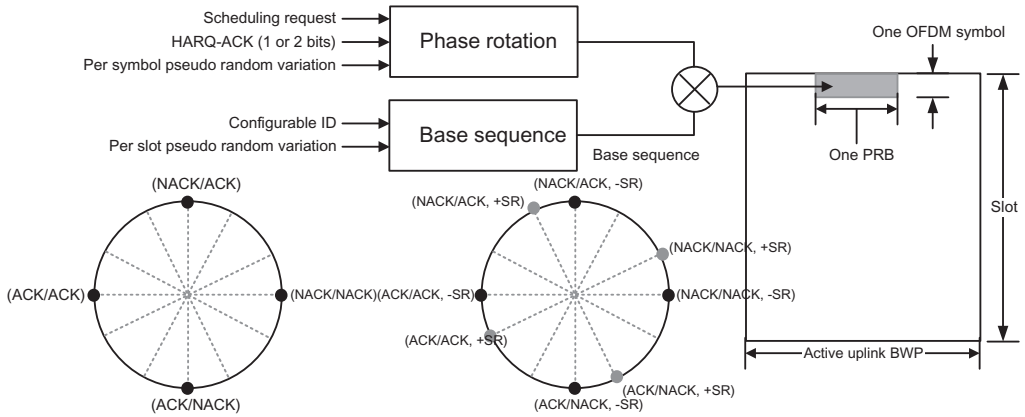


Figure 4.78
Time-frequency structure of PUCCH format 0 and signaling via phase rotation [6,14].

multiplied by amplitude scaling factor $\beta_{PUCCH-F0}$ in order to adjust the transmit power and is mapped sequentially to resource elements (k, l) assigned for PUCCH transmission in frequency-first manner on a single-antenna port [6].

As shown in Fig. 4.78, the phase rotations through parameter α represent different information bits that are separated with π and $\pi/2$ for 1- and 2-bit HARQ-ACK, respectively. In the case of a simultaneous SR, the phase rotation is increased by $\pi/4$ for 1-bit acknowledgments and by $\pi/6$ for 2-bit acknowledgments. The base sequences are configured per cell using an identity provided as part of the SI. Furthermore, a sequence hopping, where the base sequence varies on a slot-by-slot basis, can be used to randomize the interference between different cells. PUCCH format 0 is typically transmitted at the end of a slot. If two OFDM symbols are used for PUCCH format 0, the same information is transmitted on both OFDM symbols. However, the reference phase rotation as well as the frequency-domain resources may vary between the symbols, effectively creating a frequency-hopping effect [6,14].

4.2.3.1.2 PUCCH Format 1 Structure and Physical Processing

PUCCH format 1 is the long-format version of PUCCH format 0, which can carry up to 2 bits, using 4−14 OFDM symbols over one resource block per symbol in frequency. The OFDM symbols are split between symbols used for control information and symbols designated to the reference signals to enable coherent reception. The number of symbols used for control information and reference signals is a trade-off between channel estimation accuracy and energy in the information part. It was shown that a reasonable trade-off would be achieved if half of the symbols are used for reference signals. The block of bits $b(0), \ldots, b(N_{UCI} - 1) | N_{UCI} = 0, 1$ is modulated using BPSK for a single-bit payload and

QPSK for double-bit payload, resulting in a complex-valued symbol $d(0)$. The complex-valued symbol $d(0)$ is multiplied by a sequence $r_{u,v}^{(\alpha,\delta)}(n)$ such that $y(n) = d(0)r_{u,v}^{(\alpha,\delta)}(n); n = 0, 1, \ldots, 11$. The block of complex-valued symbols $y(0), \ldots, y(11)$ is block-wise spread with the (DFT) orthogonal sequence $w_i(m) = \exp(j2\pi\phi(m)/N_{SF})$ such that

$$z\left(12m'N_{SF,0}^{PUCCH-F1} + 12m + n\right) = w_i(m)y(n); \qquad n = 0, 1, \ldots, 11; m = 0, 1, \ldots, N_{SF,m'}^{PUCCH-F1} - 1$$

where $m' = 0$ when there is no intra-slot frequency hopping and $m' = 0, 1$ when intra-slot hopping is enabled. The parameter $N_{SF,m'}^{PUCCH-F1}$ and the orthogonal sequence $w_i(m)$ are defined in [6]. In case of a PUCCH transmission spanning multiple slots, the complex-valued symbol $d(0)$ is repeated for the subsequent slots. The sequence $z(n)$ is scaled with the scaling factor $\beta_{PUCCH-F1}$ in order to adjust the transmit power and is mapped sequentially to resource elements (k, l) if they are not used by DM-RS. The mapping to resource elements (k, l) designated to PUCCH transmission is in increasing order of frequency and then time over the assigned physical resource block on a single-antenna port. The time-frequency structure of PUCCH format 1 is illustrated in Fig. 4.79.

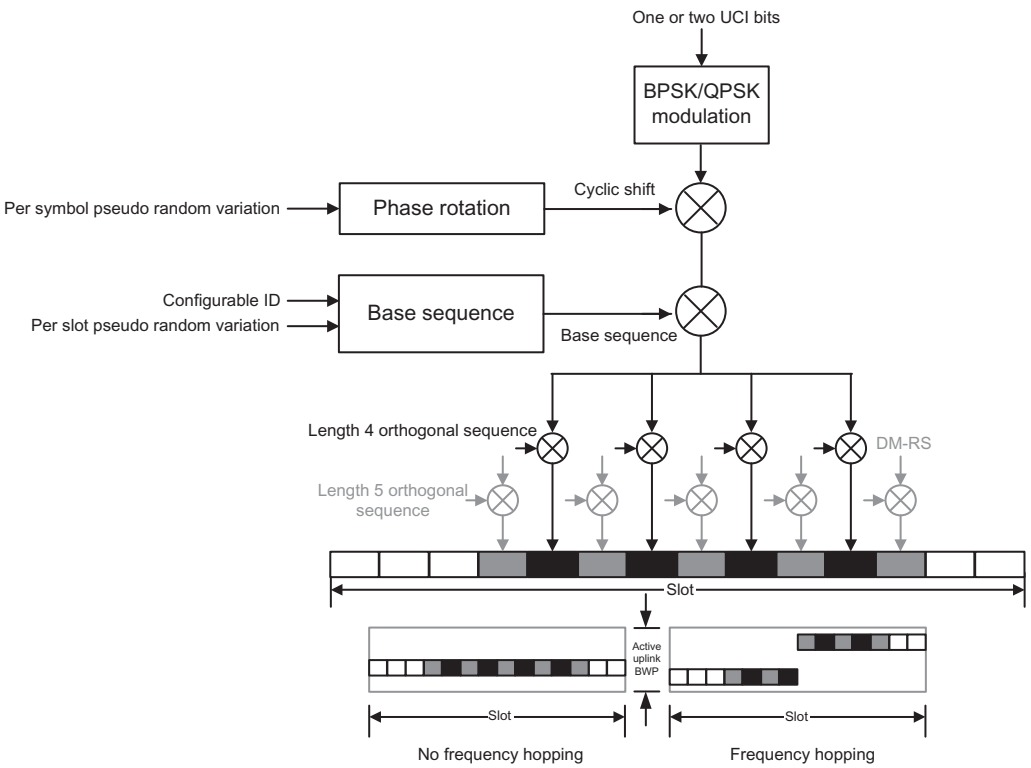


Figure 4.79
Time-frequency structure of PUCCH format 1 [6].

Unlike LTE, where PUCCH frequency-hopping was always done at the slot boundary, the NR provides additional flexibility by allowing variable PUCCH duration depending on the scheduling decisions and overall system configuration. Furthermore, since the devices are supposed to only transmit within their active bandwidth part, hopping is typically not between the edges of the overall carrier bandwidth as in LTE. Therefore, frequency hopping is configurable and determined as part of PUCCH resource configuration. The position of the hop is obtained from the length of PUCCH. If frequency hopping is enabled, one orthogonal block-spreading sequence is used per hop.

4.2.3.1.3 PUCCH Format 2 Structure and Physical Processing

PUCCH format 2 is a short PUCCH format which is used to carry more than two uplink control bits, for example, CSI report and HARQ acknowledgments, or HARQ acknowledgments per se. An SR can also be included in the bits and jointly encoded. If the payload to be transmitted by PUCCH format 2 is too large, the CSI reports are not transmitted in order to make room for more important HARQ acknowledgment bits. The overall PUCCH format 2 physical layer processing is depicted in Fig. 4.80. A CRC is added for large payload sizes

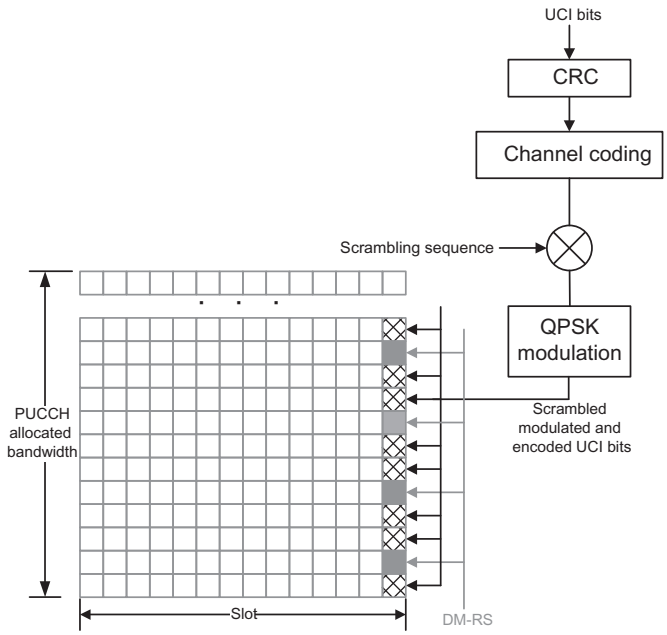


Figure 4.80
Processing and mapping of PUCCH format 2 [6].

and the resulting block of bits are encoded using Reed–Muller codes[20] for payload sizes up to 12 bits or polar codes for larger payloads. The encoded block is scrambled and QPSK modulated. The scrambling sequence is based on the device identity (C-RNTI) together with the physical-layer cell identity (or a configurable virtual cell identity), ensuring interference randomization across cells and devices using the same set of time-frequency resources. The QPSK-modulated symbols are then mapped to subcarriers across multiple resource blocks using one or two OFDM symbols. A pseudo-random QPSK sequence, mapped to every third subcarrier on each OFDM symbol, is used as a DM-RS to facilitate coherent detection at the base station [14].

The physical layer processing of UCI starts with CRC attachment and channel coding (see Fig. 4.80). Let us $a_0, a_1, \ldots, a_{N_{UCI}-1}$ denote the input UCI bit sequence, in which N_{UCI} is the payload size. If $N_{UCI} \geq 12$ the UCI bits are encoded with the polar codes and if $N_{UCI} \leq 11$, Reed–Muller codes, simplex code ($N_{UCI} = 2$), or repetition coding ($N_{UCI} = 1$) would be used to encode the UCI bits [7]. If the payload size $N_{UCI} \geq 12$, code block segmentation and CRC attachment may be performed prior to channel coding. If $N_{UCI} \geq 360, E \geq 1088$ bits or if $N_{UCI} \geq 1013, I_{seg} = 1$ segmentation is performed; otherwise, $I_{seg} = 0$ and no segmentation is done, where E is the rate-matched output sequence length. If $12 \leq N_{UCI} \leq 19$, the parity bits $p_{r0}, p_{r1}, \ldots, p_{r(N_{CRC}-1)}$ are computed by setting $N_{CRC} = 6$ bits using the generator polynomial $g_{CRC6}(D)$. The resulting sequence is $c_{l0}, c_{l1}, \ldots, c_{l(K_l-1)}$ where l is the code block number and K_l is the number of bits associated with the lth code block. If $N_{UCI} \geq 20$, the parity bits are computed by setting $N_{CRC} = 11$ bits using the generator polynomial $g_{CRC11}(D)$ [7]. Note that if the payload size $N_{UCI} \leq 11$, no CRC bits are attached.

The information bits are later encoded by channel coding block. The total number of code blocks is denoted by C and each code block is individually encoded. If $18 \leq K_l \leq 25$, the information bits are encoded with polar encoder by setting the parameters as follows: $n_{\max} = 10$, $I_{IL} = 0$, $n_{PC} = 3$, $n_{PC}^{wm} = 1$, if $E_l - K_l + 3 > 192$; otherwise, $n_{PC}^{wm} = 0$, in which E_l is the rate-matched output sequence length of the lth code block. If $K_l > 30$, the

[20] Reed–Muller codes are a family of linear error-correcting codes used in communication systems. The special cases of Reed–Muller codes include Hadamard codes, Walsh–Hadamard codes, and Reed–Solomon codes. Reed–Muller codes are denoted by $RM(d, r)$ notation, where d is the order of the code, and r determines the length of code $n = 2^r$. Reed–Muller codes are related to binary functions on field $GF(2^r)$ over the elements $\{0, 1\}$. It can be shown that $RM(0, r)$ codes are repetition codes of length $n = 2^r$, rate $R = 1/n$, and minimum distance $d_{\min} = n$, and $RM(1, r)$ codes are parity check codes of length $n = 2^r$, rate $R = (r + 1)/n$, and minimum distance $d_{\min} = n/2$. Reed–Muller codes have the following properties [15]:

1. The set of all possible exterior products of up to $d$of v_i form a basis for $\mathbf{F}_2^n$.
2. The rank of $RM(d, r)$ code is defined as $\sum_{s=0}^{r} \binom{d}{s}$.
3. $RM(d, r) = RM(d, r-1)|RM(d-1, r-1)$, where $|'$ denotes the bar product of two codes.
4. $RM(d, r)$ has minimum Hamming weight 2^{d-r}.

Table 4.26: Rate-matched UCI output sequence length E_{UCI} [7].

UCI(s) for Transmission on a PUCCH	UCI for Encoding	Value of E_{UCI}
HARQ-ACK	HARQ-ACK	$E_{UCI} = E_T$
HARQ-ACK, SR	HARQ-ACK, SR	$E_{UCI} = E_T$
CSI (consisting of one part)	CSI	$E_{UCI} = E_T$
HARQ-ACK, CSI (consisting of one part)	HARQ-ACK, CSI	$E_{UCI} = E_T$
HARQ-ACK, SR, CSI (CSI not of two parts)	HARQ-ACK, SR, CSI	$E_{UCI} = E_T$
CSI (consisting of two parts)	CSI Part 1	$E_{UCI} = \min\left(E_T, \left\lceil (O^{CSI\text{-}part1} + N_{CRC})/R_{UCI}^{\max}/Q_m \right\rceil Q_m\right)$
	CSI Part 2	$E_{UCI} = E_T - \min\left(E_T, \left\lceil (O^{CSI\text{-}part1} + N_{CRC})/R_{UCI}^{\max}/Q_m \right\rceil Q_m\right)$
HARQ-ACK, CSI (consisting of two parts)	HARQ-ACK, CSI Part 1	$E_{UCI} = \min\left(E_T, \left\lceil (O^{ACK} + O^{CSI\text{-}part1} + N_{CRC})/R_{UCI}^{\max}/Q_m \right\rceil Q_m\right)$
	CSI Part 2	$E_{UCI} = E_T - \min\left(E_T, \left\lceil (O^{ACK} + O^{CSI\text{-}part1} + N_{CRC})/R_{UCI}^{\max}/Q_m \right\rceil Q_m\right)$
HARQ-ACK, SR, CSI (consisting of two parts)	HARQ-ACK, SR, CSI Part 1	$E_{UCI} = \min\left(E_T, \left\lceil (O^{ACK} + O^{SR} + O^{CSI\text{-}part1} + N_{CRC})/R_{UCI}^{\max}/Q_m \right\rceil Q_m\right)$
	CSI Part 2	$E_{UCI} = E_T - \min\left(E_T, \left\lceil (O^{ACK} + O^{SR} + O^{CSI\text{-}part1} + N_{CRC})/R_{UCI}^{\max}/Q_m \right\rceil Q_m\right)$

information bits are encoded with polar encoder by setting the parameters as follows: $n_{\max} = 10$, $I_{IL} = 0$, $n_{PC} = 3$, $n_{PC}^{wm} = 0$. The output bits following the encoding are denoted by $d_0, d_1, \ldots, d_{N-1}$, where N is the number of coded bits.

For PUCCH format 2/3/4, the total rate-matched output sequence length E_T is given by Table 4.26, where $N_{symb,UCI}^{PUCCH,2}$, $N_{symb,UCI}^{PUCCH,3}$, *and* $N_{symb,UCI}^{PUCCH,4}$ denote the number of symbols carrying UCI for PUCCH format 2/3/4; $N_{PRB}^{PUCCH,2}$, $N_{PRB}^{PUCCH,3}$ are the number of PRBs that are determined by the UE for PUCCH format 2/3 transmission; and $N_{SF}^{PUCCH,4}$ is the spreading factor for PUCCH format 4 [7,8]. The rate matching is performed by setting $I_{BIL} = 1$ and the rate matching output sequence length to $E_l = \lfloor E_{UCI}/C_{UCI} \rfloor$, where C_{UCI} is the number of code blocks for UCI and the value of E_{UCI} is given by Table 4.26. In this table the following parameters have been used [7]:

- O^{ACK} is the number of bits for HARQ-ACK for transmission on the current PUCCH.
- O^{SR} denotes the number of bits for SR for transmission on the current PUCCH.
- $O^{CSI\text{-}part\ 1}$ is the number of bits for CSI part 1 for transmission on the current PUCCH.
- $O^{CSI\text{-}part\ 2}$ denotes the number of bits for CSI part 2 for transmission on the current PUCCH.
- $R_{UCI}^{\max}$ is the configured PUCCH maximum coding rate.

The output bit sequence after rate matching is denoted by $f_{l0}, f_{l1}, \ldots, f_{l(E_l-1)}$ where E_l is the length of rate-matched output sequence in lth code block number.

The encoded UCI payload $b(0), \ldots, b(N_E - 1)$, where N_E denotes the number of encoded bits in the payload transmitted on the physical uplink control channel, is scrambled prior to modulation, resulting in a block of scrambled bits $\tilde{b}(0), \ldots, \tilde{b}(N_E - 1)$ such that $\tilde{b}(i) = [b(i) + c(i)] \bmod 2$, in which the scrambling sequence $c(i)$ is a pseudo-random sequence that is initialized with the value $c_{init} = n_{RNTI} 2^{15} + n_{ID}$ that is derived from RRC configured parameter $n_{ID} \in \{0, 1, \ldots, 1023\}$; otherwise, $n_{ID} = N_{ID}^{cell}$. The block of scrambled bits $\tilde{b}(0), \ldots, \tilde{b}(N_E - 1)$ is QPSK modulated, resulting in a block of complex-valued modulation symbols $d(0), \ldots, d(N_E/2 - 1)$. The block of modulation symbols is scaled with the scaling factor $\beta_{PUCCH\text{-}F2}$ to adjust the transmit power and is sequentially mapped, starting with $d(0)$, to resource elements (k, l) which are reserved for PUCCH transmission and are not used by the associated DM-RS. The mapping to resource elements (k, l) is in increasing order of the frequency index k followed by time index l on a single-antenna port. PUCCH format 2 is typically transmitted at the end of a slot as illustrated in Fig. 4.80; however, it is also possible to transmit PUCCH format 2 in other positions within a slot [6].

4.2.3.1.4 PUCCH Formats 3 and 4 Structure and Physical Processing

PUCCH format 3 is the long PUCCH counterpart to PUCCH format 2, wherein more than two UCI bits can be transmitted over 4−14 OFDM symbols, where there can be multiple resource blocks on each symbol. As a result, it is the PUCCH format with the largest payload capacity. The OFDM symbols are used for carrying the UCI and the PUCCH DM-RS. The control information is encoded using Reed−Muller codes for payload sizes less than 11 bits and polar codes for larger payloads and then scrambled and modulated. The scrambling sequence is based on the UE identity (C-RNTI) together with the physical-layer cell identity (or a configurable virtual cell identity), ensuring interference randomization across cells and devices that use the same set of time-frequency resources. Prior to channel coding stage, a CRC is attached to the control information for large payloads. The encoded bits are QPSK modulated; however, there is an option to use $\pi/2$-BPSK modulation to lower the cubic metric at the expense of some loss in link performance. The complex-valued modulation symbols are distributed among the OFDM symbols and DFT precoding is performed to reduce the cubic metric and improve the power amplifier efficiency.

The structure of PUCCH format 4 is similar to that of PUCCH format 3 with the possibility to code-multiplex multiple devices over the same resources using one resource block in the frequency domain. Each OFDM symbol carries $12/N_{SF}$ unique modulation symbols. Prior to DFT-precoding, each modulation symbol is block-spread with an orthogonal sequence of length N_{SF}. The spreading factors of length two and four are supported, implying that the multiplexing capacity would be two or four devices on the same set of resource blocks.

As shown in Fig. 4.81, frequency hopping can be configured for PUCCH format 3/4 to exploit frequency diversity; however, these PUCCH formats can operate without frequency hopping. The location of the reference signal symbols depends on the frequency hopping and the length

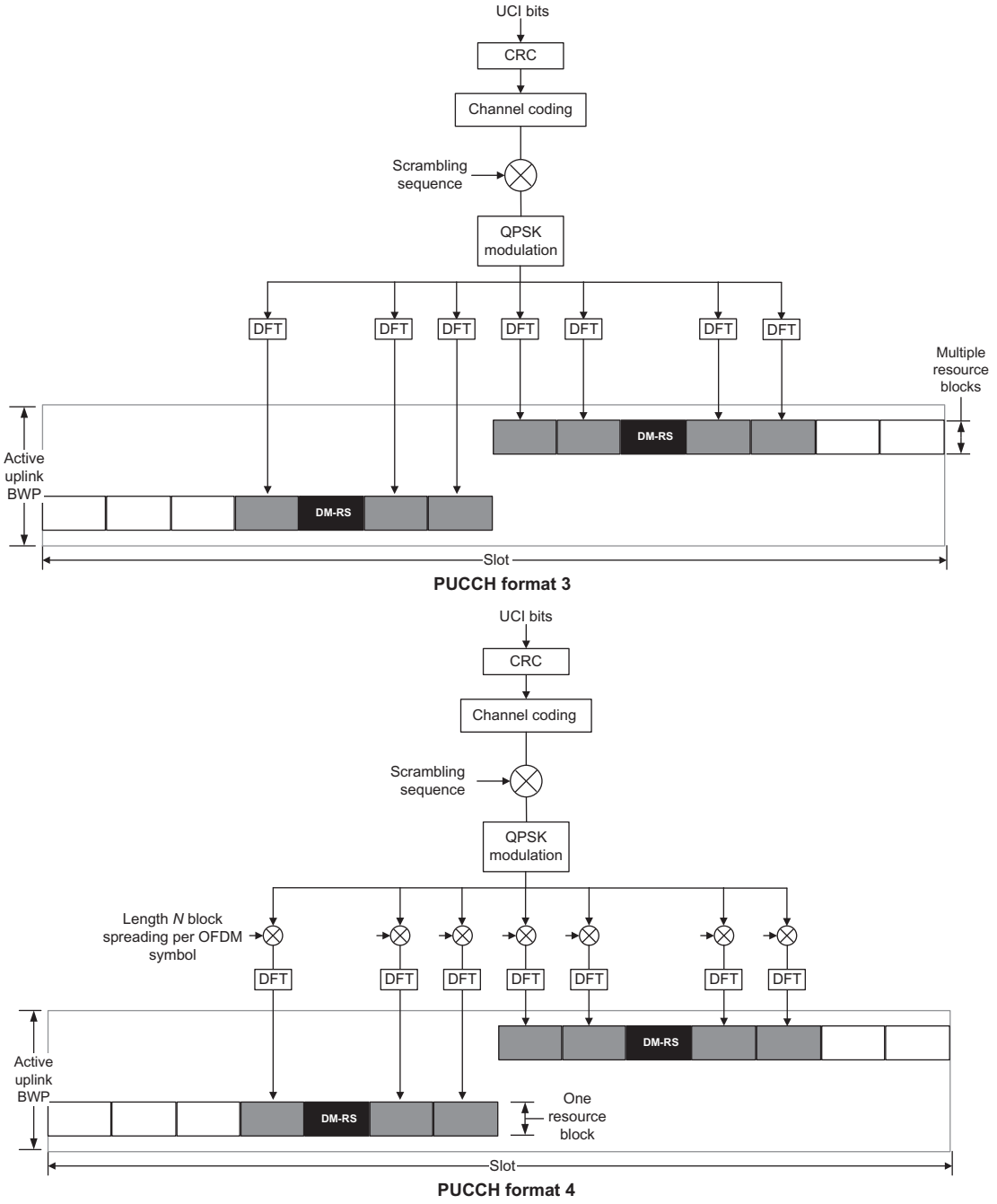


Figure 4.81
Short-PUCCH time-frequency structure [6,14].

of the PUCCH transmission, since at least one resource signal per hop is required. There is also a possibility to configure additional reference signal locations for longer PUCCH durations to ensure two reference signal instances per hop. The UCI mapping is such that higher priority is given to HARQ-ACK, SR, and CSI part 1, which are jointly encoded and mapped around the DM-RS locations, and lower priority to the remaining bits which are mapped to the remaining positions [14].

More specifically, the encoded payload bits $b(0),\ldots,b(N_E-1)$ are scrambled prior to modulation, resulting in a block of scrambled bits $\tilde{b}(0),\ldots,\tilde{b}(N_E-1)$ such that $\tilde{b}(i)=(b(i)+c(i))\text{mod }2$ where the scrambling sequence $c(i)$ is a pseudo-random sequence that is initialized with $c_{init}=n_{RNTI}2^{15}+n_{ID}$ where parameter $n_{ID}\in\{0,1,\ldots,1023\}$ is determined via RRC signaling; otherwise, it is set to the default value $n_{ID}=N_{ID}^{cell}$. The block of scrambled bits $\tilde{b}(0),\ldots,\tilde{b}(N_E-1)$ is QPSK or $\pi/2$-BPSK modulated, resulting in a block of complex-valued modulation symbols $d(0),\ldots,d(N_E/q-1)$ where $q=2$ for QPSK and $q=1$ for $\pi/2$-BPSK. For PUCCH formats 3 and 4, $M_{sc}^{PUCCH,s}=12M_{RB}^{PUCCH,s}$, in which $M_{RB}^{PUCCH,s}$ denotes the bandwidth of the PUCCH in terms of the number of resource blocks. The parameter $M_{RB}^{PUCCH,s}=2^{\alpha_2}3^{\alpha_3}5^{\alpha_5}$ for PUCCH format 3 and $M_{RB}^{PUCCH,s}=1$ for PUCCH format 4, wherein $\alpha_2,\alpha_3,\alpha_5$ is a set of non-negative integers, and $s\in\{3,4\}$ identifies the format. For PUCCH format 3, no block-wise spreading is applied and $y(lM_{sc}^{PUCCH,3}+k)=d(lm_{sc}^{PUCCH,3}+k);\quad k=0,1,\ldots,M_{sc}^{PUCCH,3}-1;\ l=0,1\ldots,(M_{symb}/M_{sc}^{PUCCH,3})-1$ where $M_{RB}^{PUCCH,3}\geq 1$ and $N_{SF}^{PUCCH,3}=1$. For PUCCH format 4, block-wise spreading is applied such that $y\left(lM_{sc}^{PUCCH,3}+k\right)=w_m(k)d\left(l\left(M_{sc}^{PUCCH,4}/N_{SF}^{PUCCH,4}\right)+k\bmod\left(M_{sc}^{PUCCH,4}/\quad N_{SF}^{PUCCH,4}\right)\right)$ where $k=0,1,\ldots,M_{sc}^{PUCCH,4}-1$ and $l=0,1,\ldots,(N_{SF}^{PUCCH,4}M_{symb}/\quad M_{sc}^{PUCCH,4})-1$. Furthermore, $M_{RB}^{PUCCH,4}=1$, $N_{SF}^{PUCCH,4}\in\{2,4\}$ and the spreading codes $w_m(k)$ are given in [6,8].

The block of complex-valued symbols $y(0),\ldots,y(N_{SF}^{PUCCH,s}M_{symb}-1)$ is DFT transformed (precoded) as follows [6]:

$$z\left(lM_{sc}^{PUCCH,s}+k\right)=\frac{1}{\sqrt{M_{sc}^{PUCCH,s}}}\sum_{m=0}^{M_{sc}^{PUCCH,s}-1}y\left(lM_{sc}^{PUCCH,s}+m\right)e^{-j\left(2\pi mk/M_{sc}^{PUCCH,s}\right)}$$

$$k=0,\ldots,M_{sc}^{PUCCH,s}-1;\quad l=0,\ldots,\left(\frac{N_{SF}^{PUCCH,s}M_{symb}}{M_{sc}^{PUCCH,s}}\right)-1$$

resulting in a block of complex-valued symbols $z(0),\ldots,z(M_{SF}^{PUCCH,s}M_{symb}-1)$, which are scaled with a scaling factor $\beta_{PUCCH,s}$ to properly adjust the transmit power and are subsequently mapped starting with $z(0)$ to resource elements (k,l) which are designated to PUCCH transmission and are not used by the associated PUCCH DM-RS. The mapping to

resource elements is done in frequency-first manner on a single-antenna port. In case of intra-slot frequency hopping, $\left\lfloor N_{symb}^{PUCCH,s}/2 \right\rfloor$ OFDM symbols are transmitted in the first hop and the remaining $N_{symb}^{PUCCH,s} - \left\lfloor N_{symb}^{PUCCH,s}/2 \right\rfloor$ symbols in the second hop where $N_{symb}^{PUCCH,s}$ is the total number of OFDM symbols used in one slot for PUCCH transmission [6]. The physical processing and structure of PUCCH formats 3 and 4 are illustrated in Fig. 4.81.

4.2.3.2 Physical Random-Access Channel

The NR supports a four-step random-access procedure similar to LTE. However, beamforming and beam tracking aspects introduced in NR random-access procedure make the overall process different from that of LTE in frequencies above 6 GHz. The UEs need to detect and select the best beam for RACH process (beam selection process) prior to PRACH sequence selection and transmission. The PRACH design in NR relies on Zadoff−Chu sequences for the preamble construction. There are three PRACH preamble formats with long sequence length of 839, two of which, with subcarrier spacing of 1.25 kHz (same as LTE), are used for LTE refarming and large cells (up to 100 km); another format with subcarrier spacing of 5 kHz is defined for high-speed scenarios (up to 500 km/h) and cell radius up to 14 km (see Tables 4.27 and 4.28). Long sequences support

Table 4.27: PRACH preamble formats for $L_{RA} = 839$ and $\Delta f_{RA} \in \{1.25, 5\}$ kHz [6].

PRACH Format	L_{RA}	Subcarrier Spacing Δf_{RA} (kHz)	Bandwidth (MHz)	N_{SEQ}	T_{SEQ}	T_{CP}	T_{GP}	Support for Restricted Sets
0	839	1.25	1.08	1	$24576T_s$	$3168T_s$	$2976T_s$	Type A, Type B
1		1.25	1.08	2	$24576T_s$	$21024T_s$	$21984T_s$	
2		1.25	1.08	4	$24576T_s$	$4688T_s$	$29264T_s$	
3		5	4.32	1	$24576T_s$	$3168T_s$	$2976T_s$	

Table 4.28: Preamble formats for $L_{RA} = 139$ and $\Delta f_{RA} = 2^{\mu} \times 15$ kHz where $\mu \in \{0, 1, 2, 3\}$ and $\kappa = 64$ [6].

PRACH Format	L_{RA} Samples	Δf_{RA} (kHz)	T_{SEQ}	N_{SEQ}	N_u^{RA} Samples	N_{CP}^{RA} Samples
A1	139	$2^{\mu} \times 15$	$2048T_s$	2	$2 \times 2048\kappa 2^{-\mu}$	$288\kappa 2^{-\mu}$
A2				4	$4 \times 2048\kappa 2^{-\mu}$	$576\kappa 2^{-\mu}$
A3				6	$6 \times 2048\kappa 2^{-\mu}$	$864\kappa 2^{-\mu}$
B1				2	$2 \times 2048\kappa 2^{-\mu}$	$216\kappa 2^{-\mu}$
B2				4	$4 \times 2048\kappa 2^{-\mu}$	$360\kappa 2^{-\mu}$
B3				6	$6 \times 2048\kappa 2^{-\mu}$	$504\kappa 2^{-\mu}$
B4				12	$12 \times 2048\kappa 2^{-\mu}$	$936\kappa 2^{-\mu}$
C0				1	$1 \times 2048\kappa 2^{-\mu}$	$1240\kappa 2^{-\mu}$
C2				4	$4 \times 2048\kappa 2^{-\mu}$	$2048\kappa 2^{-\mu}$

unrestricted sets and restricted sets of Type A and Type B, while short sequences support unrestricted sets only. Considering network beam-sweeping reception within a RACH occasion, NR introduced a new set of PRACH preamble formats of shorter sequence length of 139 on 1, 2, 4, 6, and 12 OFDM symbols and subcarrier spacings of 15, 30, 60, and 120 kHz. The new formats are composed of single or consecutive repeated RACH sequences. The cyclic prefix is inserted at the beginning of the preambles, and the guard time is appended at the end of the preambles, while the cyclic prefix and gap between RACH sequences is omitted. For both short and long PRACH preamble sequences, the network can also conduct beam-sweeping reception between RACH occasions.

Multiple RACH preamble formats are defined with one or more PRACH symbols, and different cyclic prefix and guard time lengths. The PRACH preamble configuration is signaled to the UE in the SI. The UE calculates the PRACH transmit power for the retransmission of the preamble based on the most recent estimate of pathloss and power ramping counter. If the UE conducts beam switching, the counter of power ramping remains unchanged. The SI informs the UE of the association between the SS blocks and the RACH resources. The threshold of the SS block for RACH resource association is based on the RSRP and is network configurable.

Prior to initiation of the physical random-access procedure, the physical layer of the UE must receive a set of SS/PBCH block indices and provide the UE RRC sublayer with the corresponding set of RSRP measurements conducted on those SS/PBCH candidates. The information required for the UE physical layer prior to PRACH transmission includes PRACH preamble format, time resources, and frequency resources for PRACH transmission as well as the parameters for determining the root sequences and their cyclic shifts in the PRACH preamble sequence set including index to logical root sequence table, cyclic shift N_{CS}, and set type, that is, unrestricted, restricted set A, or restricted set B.

Physical random-access procedure is triggered following a request from the UE RRC sublayer or by a PDCCH command (control element) with the following parameters [8]: configuration for PRACH transmission, preamble index, preamble subcarrier spacing, transmit power P_{PRACH}^{target}, RA-RNTI, and a PRACH resource. The PRACH preamble is transmitted using the selected PRACH format with transmission power value over the designated PRACH resource.

The SS/PBCH block indices are mapped to PRACH occasions (see Fig. 4.82); first, in increasing order of preamble indices within a single PRACH occasion followed by, in increasing order of frequency resource indices of frequency-multiplexed PRACH occasions, then, in increasing order of time resource indices of time-multiplexed PRACH occasions within a PRACH slot and, finally, in increasing order of indices of PRACH slots. The association period, starting from frame 0, for the mapping of SS/PBCH blocks to PRACH occasions

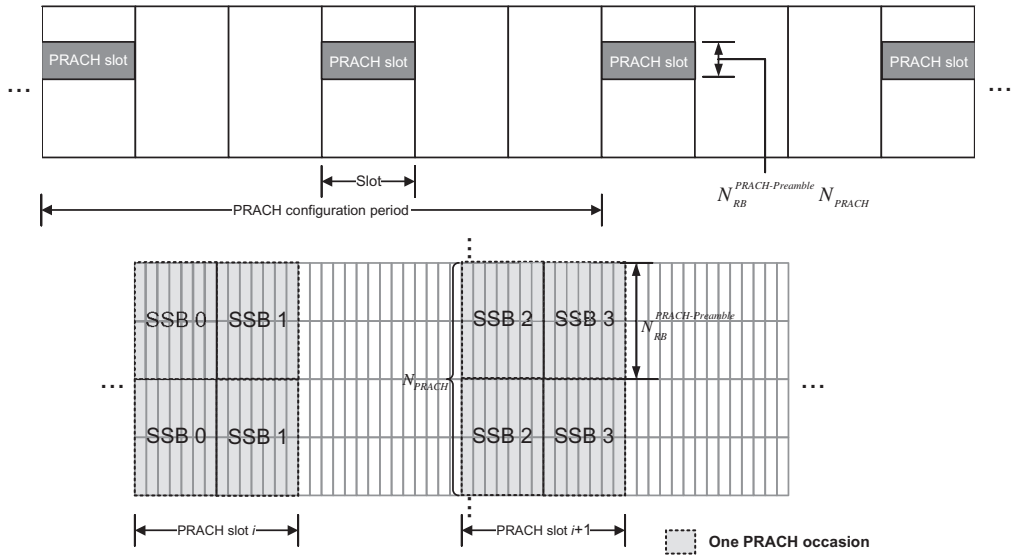


Figure 4.82
Structure of NR PRACH opportunities in time and frequency domain [14].

Table 4.29: Mapping between PRACH configuration period and SS/PBCH block to PRACH occasion association period [8].

PRACH Configuration Period (ms)	Association Period (Number of PRACH Configuration Periods)
10	{1,2,4,8,16}
20	{1,2,4,8}
40	{1,2,4}
80	{1,2}
160	{1}

is the smallest value in a set (see Table 4.29) determined by the PRACH configuration period such that N_{TX}^{SSB} SS/PBCH blocks are mapped at least once to the PRACH occasions within the association period. A UE obtains the parameter N_{TX}^{SSB} from *SystemInformationBlockType1*. If after an integer number of SS/PBCH blocks to PRACH occasions mapping cycles within the association period, there is a set of PRACH occasions that are not mapped to N_{TX}^{SSB} SS/PBCH blocks; no SS/PBCH blocks are mapped to the set of PRACH occasions. An association pattern period includes one or more association periods and is calculated such that a pattern between PRACH occasions and SS/PBCH blocks repeats at most every 160 ms. The PRACH

occasions that are not associated with SS/PBCH blocks after an integer number of association periods, if any, are not used for PRACH transmissions [8].

If a random-access procedure is initiated by a PDCCH command, the UE must transmit the PRACH preamble in the first available PRACH occasion for which the time interval between the last symbol of the PDCCH reception and the first symbol of the PRACH transmission is larger than or equal to $T_{N_2} + \Delta_{BWP\text{-}Switching} + \Delta_{Delay}$ (in milliseconds), where T_{N_2} is the equivalent time duration of N_2 symbols corresponding to PUSCH preparation time assuming certain PUSCH processing capability, the parameter $\Delta_{BWP-Switching} = 0$, if uplink active BWP does not change, and $\Delta_{Delay} = 0.5$ ms for FR1 and $\Delta_{Delay} = 0.25$ ms for FR2 [8].

The PRACH preamble transmission can occur within a configurable subset of slots known as the PRACH slots (see Fig. 4.82) that are repeated every PRACH configuration period. There may be multiple PRACH occasions within each PRACH slot in the frequency-domain that cover $N_{RB}^{PRACH\text{-}Preamble} N_{PRACH}$ consecutive resource blocks where $N_{RB}^{PRACH\text{-}Preamble}$ is the preamble bandwidth measured in number of resource blocks and N_{PRACH} is the number of frequency-domain PRACH occasions. For a given preamble type, corresponding to a certain preamble bandwidth, the overall available time-frequency PRACH resources within a cell can be described by the following parameters: a configurable PRACH periodicity that can range from 10 to 160 ms; a configurable set of PRACH slots within the PRACH period; a configurable frequency-domain PRACH resource given by the index of the first resource block in the resource and the number of frequency-domain PRACH occasions [14].

In NR, the set of random-access preambles $x_{u,v}(n)$ is generated based on Zadoff−Chu sequences such that $x_{u,v}(n) = x_u[(n + C_v) \bmod L_{RA}]$ wherein $x_u(i) = \exp(-j\pi ui(i+1)/L_{RA})$ is a Zadoff−Chu sequence of length L_{RA} and root index u; and $i = 0, 1, \ldots, L_{RA} - 1$. The frequency-domain representation of the preamble sequence is obtained by taking L_{RA} − point DFT of the sequence $x_{u,v}(n)$, resulting in $y_{u,v}(n) = \sum_{m=0}^{L_{RA}-1} x_{u,v}(m) \exp(-j2\pi mn/L_{RA})$ where the sequence length $L_{RA} = 839$ or $L_{RA} = 139$ depend on the PRACH preamble format (see Tables 4.27 and 4.28). There are 64 preambles in each time--frequency PRACH occasion, numbered in increasing order of cyclic shift C_v of a logical root sequence and increasing order of the logical root sequence index, starting with the index obtained via RRC signaling. The sequence number u is obtained from the logical root sequence index [6]. The output of the DFT is then repeated N_{SEQ} times, after which a cyclic prefix is inserted. For the PRACH preamble, the cyclic prefix is not inserted per OFDM symbol, rather it is inserted only once for the block of N_{SEQ} repeated symbols (see Fig. 4.83).

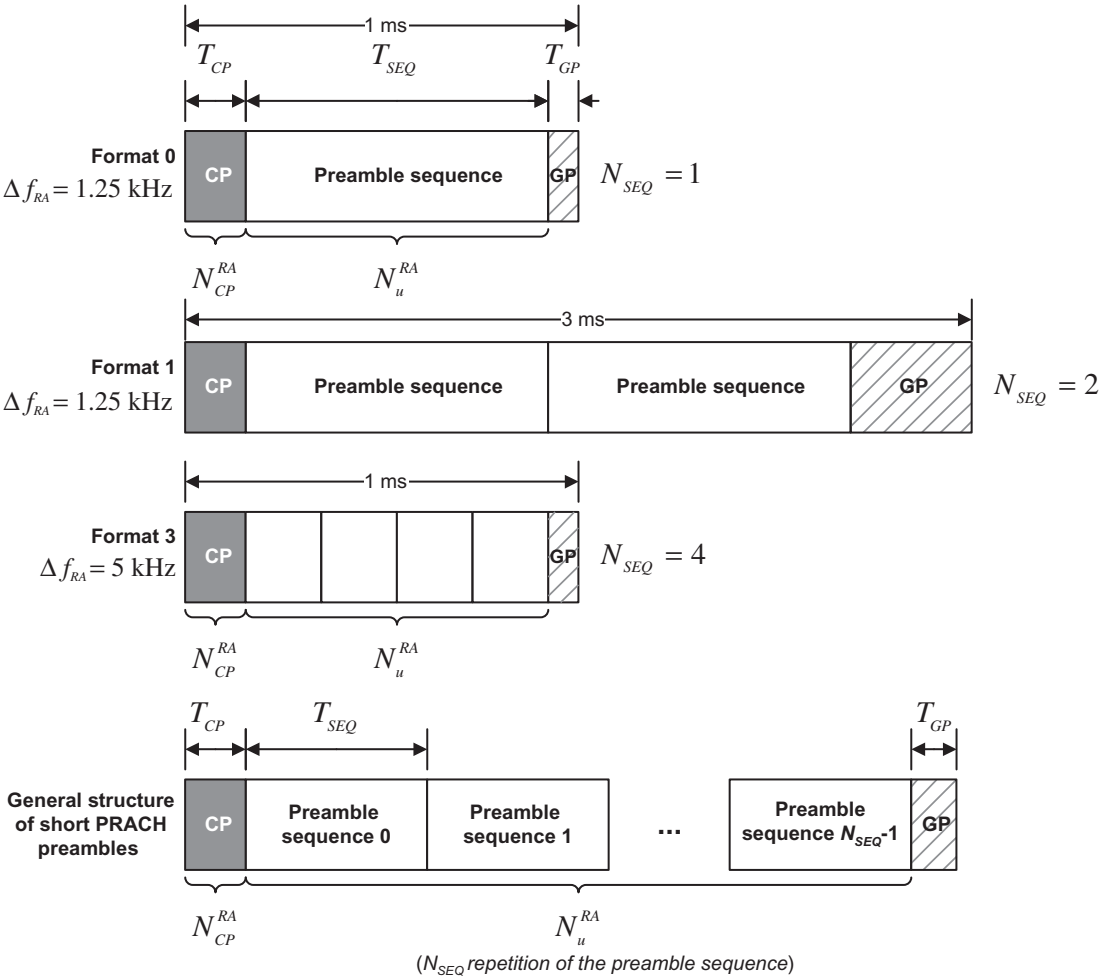


Figure 4.83
NR PRACH preamble structures [6].

The time-domain representation of PRACH signal $s_{PRACH}(t,l)$ on a single-antenna port is given as follows [6]:

$$s_{PRACH}(t,l) = \sum_{k=0}^{L_{RA}-1} a_k e^{j2\pi(k+Kk_1+\bar{k})\Delta f_{RA}(t-N_{CP,l}^{RA}T_c-t_{start}^{RA})}; \quad K = \frac{\Delta f}{\Delta fRA}$$

$$k_1 = k_o^{\mu} + N_{BWP,i}^{start}N_{sc}^{RB} + n_{RA}^{start}N_{sc}^{RB} + n_{RA}N_{RB}^{RA}N_{sc}^{RB} - \frac{N_{grid}^{size,\mu}N_{sc}^{RB}}{2}$$

$$t_{start}^{RA} \le t < t_{start}^{RA} + \left(N_u + N_{CP,l}^{RA}\right)T_c$$

$$k_0^{\mu} = \left(N_{grid}^{start}(\mu) + \frac{N_{grid}^{size}(\mu)}{2}\right)N_{sc}^{RB} - \left(N_{grid}^{start}(\mu_0) + \frac{N_{grid}^{size}(\mu_0)}{2}\right)N_{sc}^{RB}2^{\mu_0-\mu}$$

Table 4.30: Supported combinations of Δf_{RA} and Δf, and $\overline{k}$ [6].

L_{RA}	Δf_{RA} for PRACH	Δf for PUSCH	N_{RB}^{RA} (RBs) for PUSCH	$\overline{k}$
839	1.25	15	6	7
839	1.25	30	3	1
839	1.25	60	2	133
839	5	15	24	12
839	5	30	12	10
839	5	60	6	7
139	15	15	12	2
139	15	30	6	2
139	15	60	3	2
139	30	15	24	2
139	30	30	12	2
139	30	60	6	2
139	60	60	12	2
139	60	120	6	2
139	120	60	24	2
139	120	120	12	2

In the preceding expression

- $\overline{k}$ is given in Table 4.30 [6].
- Δf is the subcarrier spacing of the active uplink bandwidth part during the initial access; otherwise, Δf is the subcarrier spacing of the active uplink bandwidth part (see Table 4.30 for permissible values).
- $N_{BWP,i}^{start}$ is the lowest numbered resource block of the initial active uplink bandwidth part based on common resource block indexing and is derived via RRC parameter *initialUplinkBWP* during initial access; otherwise, $N_{BWP,i}^{start}$ is the lowest numbered resource block of the active uplink bandwidth part based on common resource block indexing and is derived by the higher layer parameter *BWP-Uplink*.
- n_{RA}^{start} is the frequency offset of the lowest PRACH transmission occasion in the frequency-domain relative to *PRB_0* of the initial active uplink bandwidth part given by the RRC parameter *msg1-FrequencyStart* during initial access associated with the initial active uplink bandwidth part; otherwise, n_{RA}^{start} is the frequency offset of lowest PRACH transmission occasion in frequency domain with respect to physical resource block 0 of the active uplink bandwidth part given by the RRC parameter *prach-frequency-start* associated with the active uplink bandwidth part.
- n_{RA} is the PRACH transmission occasion index in the frequency-domain for a given PRACH transmission occasion in time.
- N_{RB}^{RA} is the number of resource blocks that are occupied by PRACH preamble.

The starting position of PRACH preamble in a subframe t_{start}^{RA} when $\Delta f_{RA} \in \{1.25, 5, 15, 30\}$ kHz or in a slot with 60 kHz subcarrier spacing when

$\Delta f_{RA} \in \{60, 120\}$ kHz is defined as $t_{start}^{RA} = 0$ when $l = 0$; otherwise, $t_{start}^{RA} = t_{start,l-1}^{\mu} + \left(N_u^{\mu} + N_{CP,l-1}^{\mu}\right) T_c$. The subframe or 60 kHz slot is assumed to start at $t = 0$. The timing advance is assumed to be zero $N_{TA} = 0$. The numerology corresponding to $\Delta f_{RA} \in \{1.25, 5\}$ kHz is assumed to be $\mu = 0$; otherwise, it is given by $\Delta f_{RA} \in \{15, 30, 60, 120\}$ kHz, and the symbol position l is given by $l = l_0 + n_t^{RA} n_{duration}^{RA} + 14 n_{slot}^{RA}$ where l_0 is the starting symbol, n_t^{RA} is the PRACH transmission occasion within the PRACH slot, numbered in increasing order from 0 to $N_t^{RA,slot} - 1$ within a RACH slot, $N_{duration}^{RA}$ is given in [6], and the n_{slot}^{RA} depends on Δf_{RA}, that is, if $\Delta f_{RA} \in \{1.25, 5, 15, 60\}$ kHz then $n_{slot}^{RA} = 0$; otherwise if $\Delta f_{RA} \in \{30, 120\}$ kHz and either of the *number of PRACH slots within a subframe* or *number of PRACH slots within a 60 kHz slot* is equal to 1, then $n_{slot}^{RA} = 1$; otherwise, $n_{slot}^{RA} = 0, 1$. The quantities L_{RA} and N_u are the length of the PRACH sequence and the number of samples in a PRACH symbol, and $N_{CP,l}^{RA} = N_{CP}^{RA} + 16\kappa n$ wherein $n = 0$ for $\Delta f_{RA} \in \{1.25, 5\}$ kHz. For $\Delta f_{RA} \in \{15, 30, 60, 120\}$ kHz, n is the number of times the interval $[t_{start}^{RA}, t_{start}^{RA} + (N_u^{RA} + N_{CPt}^{RA})T_c]$ overlaps with either time instance zero or time instance $(\Delta f_{\max} N_f / 2000) T_c = 0.5$ ms in a subframe [6].

The parameters *ZeroCorrelationZoneConfig* and *prach-RootSequenceIndex* (defined in [6]) are used to generate the random-access signatures for each cell, which are required to be distinct across neighboring cells. There is a relationship between the preamble format and the cell radius, which means that the selection of *ZeroCorrelationZoneConfig* parameter is related to the cell radius. The parameters *ZeroCorrelationZoneConfig* and *prach-RootSequenceIndex* are derived from *SystemInformationBlockType1*. The random-access sequences are generated via selection of a Zadoff−Chu sequence (1 out of 839 or 139) given by *prach-RootSequenceIndex* and a cyclic shift that is used 64 times to generate the 64 random-access signatures from the Zadoff−Chu sequence selected. The cyclic shift is indirectly provided to the UE via the parameter *ZeroCorrelationZoneConfig*. The cyclic shift is also related to the cell size. The relationship between the cyclic shift and the cell size is given by $(N_{CS} - 1)(800\,\mu s/839) \geq \text{RTD} + \tau_{Delay\_Spread}$. If $\Delta f_{RA} = 1.25$ kHz, the PRACH symbol duration is 0.8 ms (0.133 ms in case of 139). The round-trip delay can be written as $\text{RTD} = 2R_{cell}/c$; therefore $R_{cell} \leq c[(N_{CS} - 1)(800\,\mu s/839) - \tau_{Delay\_Spread}]/2$. As an example, if we assume that *ZeroCorrelationZoneConfig* is 12 then from [6] and assuming $\Delta f_{RA} = 1.25$ kHz, $N_{CS} = 119$. Furthermore, if $\tau_{Delay\_Spread} = 6\,\mu s$ then the cell size will be approximately 15.97 km. Note that the smaller the cyclic shift, the smaller cell size. The delay spread in the preceding expression is derived empirically and the value of the delay spread is typically different for rural, suburban, urban and dense urban environments. In practice, the *ZeroCorrelationZoneConfig* parameter is a pointer to a table that provides a set of available cyclic shifts in the cell, where different tables indicated by this parameter have different distances between the cyclic shifts, thus providing larger or smaller zones or timing errors for which orthogonality or zero correlation can be maintained.

Table 4.31: Random-access configurations for TDD mode in FR1 [6].

PRACH Configuration Index	Preamble Format	n_{SFN} mod $x = y$		Subframe Number	Starting Symbol	Number of PRACH Slots Within a Subframe	$N_t^{PRACH,slot}$ Number of Time-Domain PRACH Occasions Within a RACH Slot	$N_{duration}^{RA}$ PRACH Duration
		x	y					
0	0	16	1	9	0	–	–	0
1	0	8	1	9	0	–	–	0
2	0	4	1	9	0	–	–	0
3	0	2	0	9	0	–	–	0
4	0	2	1	9	0	–	–	0
5	0	2	0	4	0	–	–	0
6	0	2	1	4	0	–	–	0
7	0	1	0	9	0	–	–	0
8	0	1	0	8	0	–	–	0
9	0	1	0	7	0	–	–	0
10	0	1	0	6	0	–	–	0
11	0	1	0	5	0	–	–	0
12	0	1	0	4	0	–	–	0
...	...	...	...	...	...	...	...	...
251	C2	1	0	3,4,8,9	2	2	2	6
252	C2	1	0	0,1,2,3,4,5,6,7,8,9	8	1	1	6
253	C2	1	0	1,3,5,7,9	2	1	2	6
254	C2	8	1	9	8	2	1	6
255	C2	4	1	9	8	1	1	6

The PRACH preamble sequence is mapped to physical resources such that $a_k^{(p,RA)} = \beta_{PRACH} y_{u,v}(k)$; $k = 0, 1, \ldots, L_{RA} - 1$ where β_{PRACH} is a transmit power adjustment scaling factor, p is the antenna port from which the PRACH is transmitted. The PRACH preambles can only be transmitted in the time resources that are signaled via RRC parameter *prach-ConfigurationIndex* and further depend on frequency range FR1 or FR2 where the system is deployed and the spectrum type. The PRACH preambles can only be transmitted in the frequency resources specified by parameter *msg1-FrequencyStart*. The PRACH frequency resources $n_{RA} \in \{0, 1, \ldots, M - 1\}$, in which the parameter M is derived from the RRC parameter *msg1-FDM*, are numbered in increasing order within the initial active uplink bandwidth part during initial access, starting from the lowest frequency. For the purpose of slot numbering, it is assumed that subcarrier spacing is 15 kHz for FR1 and 60 kHz for FR2. Table 4.31 provides random-access configurations for TDD mode in FR1 [6,8].

The transmission power for PRACH $P_{PRACH}^{klm}(i)$ on the kth uplink BWP of the lth carrier based on a certain SS/PBCH block determination for the mth serving cell in transmission period i is determined by the UE as $P_{PRACH}^{klm}(i) = \min [P_{CMAX}^{lm}(i), P_{PRACH-target}^{lm} + PL_{klm}]$ (dBm), wherein $P_{CMAX}^{lm}(i)$ is the configured UE transmission power for the lth carrier in the

mth serving cell within transmission period i, $P^{lm}_{PRACH-target}$ is the PRACH preamble target reception power *PREAMBLE_RECEIVED_TARGET_POWER* signaled via RRC for the kth uplink BWP on lth carrier in the mth serving cell, and PL_{klm} is the calculated pathloss for the kth uplink BWP corresponding to the lth carrier for the current SS/PBCH block in the mth serving cell calculated by the UE in decibels. If within a random-access response window, the UE cannot receive a random-access response that contains a preamble identifier corresponding to the preamble sequence transmitted by the UE, the UE will typically ramp up (in steps) the transmission power up to a certain limit for the subsequent PRACH transmissions. If prior to PRACH retransmission, the UE changes the spatial domain transmission filter; the physical layer will notify the higher layers to suspend the power ramping counter [8].

The random-access preamble sequence can be generated at the system sampling rate, by means of a large IDFT unit. The cyclic shift can be implemented either in the time domain after the IDFT or in the frequency domain before the IDFT through a phase shift. For all possible system sampling rates, both cyclic prefix and sequence duration correspond to an integer number of samples. The method of Fig. 4.84 does not require any time-domain filtering in the baseband but requires large IDFT sizes (up to 24576 for a 20 MHz spectrum allocation), which are practically prohibitive. Therefore, another option for generating the PRACH preamble consists of using small-sized IDFT and shifting the preamble to the required frequency location through time-domain up-sampling and filtering (hybrid frequency/time-domain generation). Assuming that the preamble sequence length is 839, the smallest IFFT size that can be used is 1024. The sizes of the random-access cyclic prefix and preamble sequence duration have been chosen to provide an integer number of samples at the system sampling rate. The cyclic prefix can be inserted before the up-sampling and time-domain frequency shift, in order to minimize the intermediate storage requirements.

Assuming sampling rate of 30.72 MHz and considering that the random-access preamble spans 0.8 ms, it can be concluded that the number of samples in time is equal to 24576. Furthermore, if the PRACH subcarrier spacing is assumed to be 1.25 kHz and the subcarrier spacing for PUSCH and PUCCH is 15 kHz, in order to maintain the same sampling rate, a 12×2048 point DFT operation would be needed for the PRACH signal generation at the transmitter side, if the entire processing is done in the frequency domain. An alternative approach is to use time-domain signal generation and extraction which involves up-sampling and filtering operations at the transmitter. The drawback of time-domain implementation is that the up-sampling from 1.08 MHz to the system sampling-rate of 30.72 MHz is difficult to implement.

The implementation of the PRACH signal at the gNB receiver can take a frequency-domain or a hybrid time/frequency-domain approach. As illustrated in Fig. 4.84 as an example, the common parts to both approaches are the cyclic prefix removal, which always occurs at the

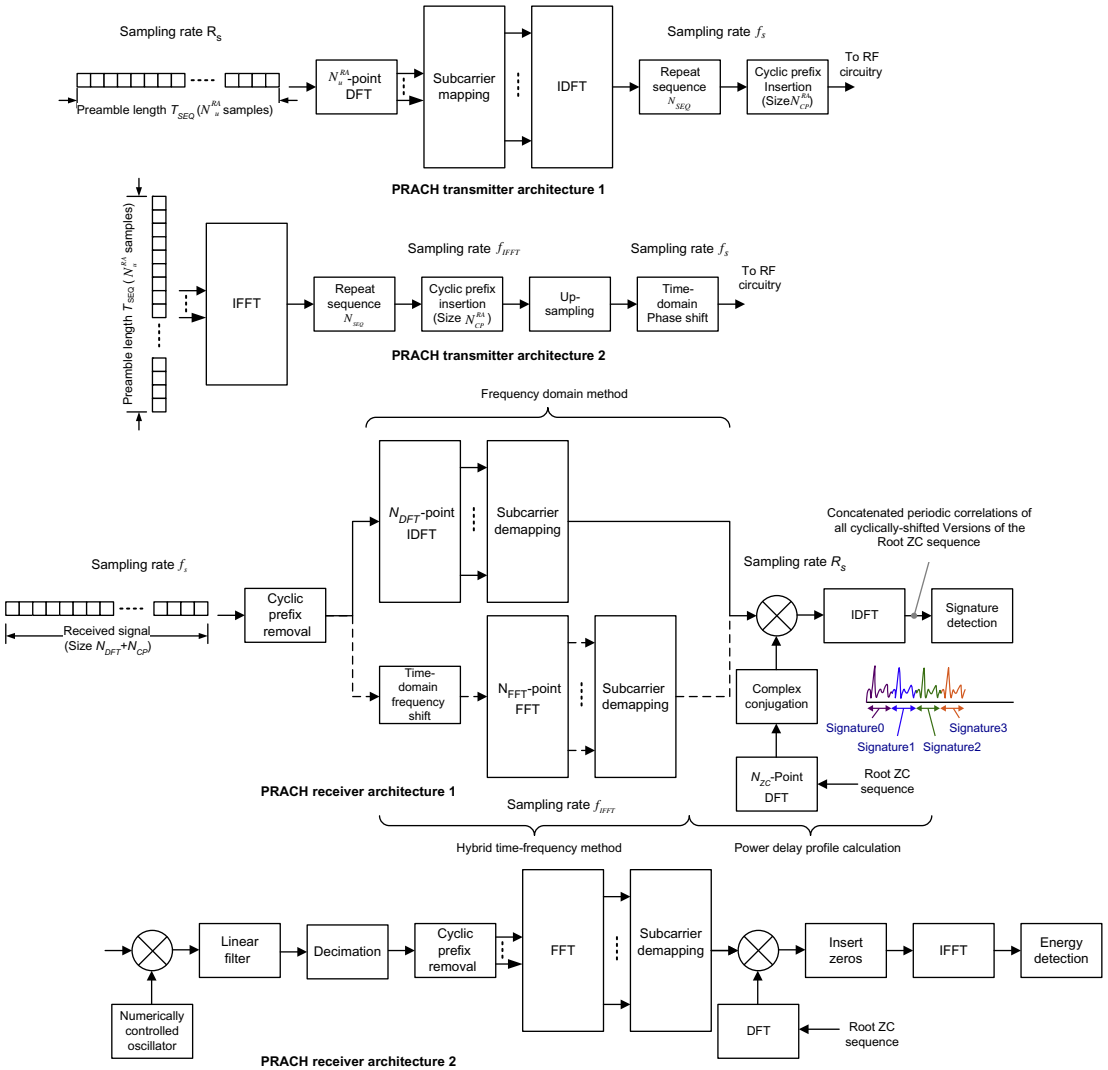


Figure 4.84
Example PRACH transmitter and receiver structure [15].

front-end at the system sampling rate, the power delay profile calculation, and signature detection. The two approaches differ only in the computation of the subcarriers carrying the PRACH signal(s). The frequency-domain method computes the full range of subcarriers used for uplink transmission over the system bandwidth from 0.8 ms-long received input samples. As a result, the PRACH subcarriers are directly extracted from the set of uplink subcarriers, which does not require any frequency shift or time-domain filtering but involves an extremely large DFT computation. Note that even though DFT size $N_{DFT} = n2^m$, and we

can use fast and effcient DFT computation algorithms, the DFT computation cannot start until the complete sequence is stored in memory, which increases the processing delay.

4.2.3.2.1 Four-Step Random-Access Procedure

From the UE physical layer perspective, the RACH procedure consists of transmission of random-access preamble (Msg1) in a PRACH occasion, receiving random-access response message via PDCCH/PDSCH (Msg2), and transmission of Msg3 in PUSCH, and receiving PDSCH for contention resolution. If the random-access procedure is initiated by a PDCCH command, the PRACH preamble is transmitted with the same subcarrier spacing. The random-access procedure comprises four steps. However, before the UE can attempt to access the network, it must synchronize to the downlink and receive the SI via PBCH and PDCCH/PDSCH. Upon receiving the SI, the UE would have the knowledge of PRACH configuration and transmission parameters such as PRACH preamble format, time-frequency resources to transmit PRACH, the parameters for determining the root sequences and their cyclic shifts in the PRACH preamble sequence set, index to the logical root sequence table, cyclic shifts, and the associated set type, that is, unrestricted, restricted Type A, or restricted Type B [6,8,12]. More specifically, the RACH procedure consists of the following 4 steps [11,12]:

In the first step, the UE transmits a PRACH preamble associated with an RA-RNTI, if all conditions for PRACH transmission are met [12]. The gNB calculates the RA-RNTI associated with the PRACH occasion, in which the random-access preamble is transmitted, as follows $RA\text{-}RNTI = 1 + s_{id} + 14t_{id} + 14 \times 80f_{id} + 14 \times 80 \times 8ul_{carrier\_id}$ where $0 \le s_{id} < 14$ is the index of the first OFDM symbol of the specified PRACH;$0 \le t_{id} < 80$ denotes the index of the first slot symbol of the specified PRACH in a system frame;$0 \le f_{id} < 8$ is the index of the specified PRACH in the frequency domain; and $ul_{carrier\_id}$ is the uplink carrier used for Msg1 transmission ($ul_{carrier\_id} = 0$: NR uplink carrier, $ul_{carrier\_id} = 1$: SUL carrier). The frequency-domain location (resource) for PRACH preamble is determined by the RRC parameter *msg1-FDM* and *msg1-FrequencyStart*. The time-domain location (resource) for PRACH preamble is determined by the RRC parameter *prach-ConfigurationIndex*.

In the second step, following the PRACH transmission, the UE awaits random-access response from the gNB which would be sent through a DCI scrambled with RA-RNTI value calculated as above. The UE attempts to detect a PDCCH with the corresponding RA-RNTI within the period of *ra-ResponseWindow*. The UE searches for the DCI in the Type 1 PDCCH common search space. The DCI format for scheduling RAR message on PDSCH is DCI format 1_0 scrambled with RA-RNTI. The resource allocation type for the Msg2 on PDSCH is resource allocation Type 1. The frequency-domain resource allocation for the PDSCH carrying RAR message is specified by DCI format 0_1. The time-domain resource allocation for the RAR message on PDSCH is specified by DCI format 1 and *PDSCH-ConfigCommon*. The RAR window is configured by *rar-WindowLength* information element

in a SIB message. If the UE successfully detects the PDCCH, it can decode PDSCH carrying the RAR message. After decoding the RAR message, the UE checks, if the random-access preamble ID (RAPID) in the RAR message matches the RAPID assigned to the UE. The PDCCH and PDSCH associated with the process are expected to use the same subcarrier spacing and cyclic prefix as SIB1. Note that the gNB is not expecting any HARQ-ACK for the RAR message. The gNB may conclude that UE has successfully received and decoded the RAR message, if the UE does not retransmit PRACH, which would happen if the UE does not detect the DCI format 1_0 with CRC scrambled with the corresponding RA-RNTI within the RAR window, or if the UE does not correctly receive the transport block in the corresponding PDSCH within that window.

In the third step, the UE must determine whether it should apply transform precoding for Msg3 on PUSCH, based on the RRC parameter *msg3-transformPrecoder*. The UE determines the subcarrier spacing for Msg3 on PUSCH based on the RRC parameter *SubcarrierSpacing* in *BWP-UplinkCommon*. The UE then transmits Msg3 on PUSCH to the same serving cell to which it had sent the PRACH.

In the fourth step, Msg4 is transmitted to the UE for contention resolution. The UE starts *ra-ContentionResolutionTimer* and monitors PDCCH with TC-RNTI while *ra-ContentionResolutionTimer* is running. The UE looks for the DCI in Type 1 PDCCH common search space. If the PDCCH is successfully detected, the UE proceed to decode PDSCH carrying the MAC control element, and at the same time, it sets the value of the C-RNTI to TC-RNTI and discards *ra-ContentionResolutionTimer*. The UE considers the RACH procedure as successfully completed. Once the UE successfully decodes Msg4 (contention resolution), it sends HARQ-ACK for the data (PDSCH carrying Msg4). In response to the PDSCH reception with the UE contention resolution identity, the UE transmits HARQ-ACK information on a PUCCH. This procedure is illustrated in Fig. 4.85.

4.2.3.2.2 Two-Step Random-Access Procedure

As we mentioned in the previous section, the NR Rel-15 supports a four-step RACH procedure. A two-step RACH procedure can be utilized, wherein the UE combines Msg1 and Msg3 of the four-step RACH procedure into one message, for example, MsgA, and transmits it to the base station. The base station also combines Msg2 and Msg4 of the four-step RACH procedure and sends it as a response, for example, MsgB, to the UE. The combining of the messages provides a low-latency RACH procedure which is useful for low-latency applications and services. More specifically, in two-step RACH, MsgA is a combination of the PRACH preamble in Msg1 and the data contained in Msg3, while MsgB combines the random-access response in Msg2 and the contention resolution information of Msg4. Fig. 4.86 shows and compares the two-step and four-step contention-based random-access procedures.

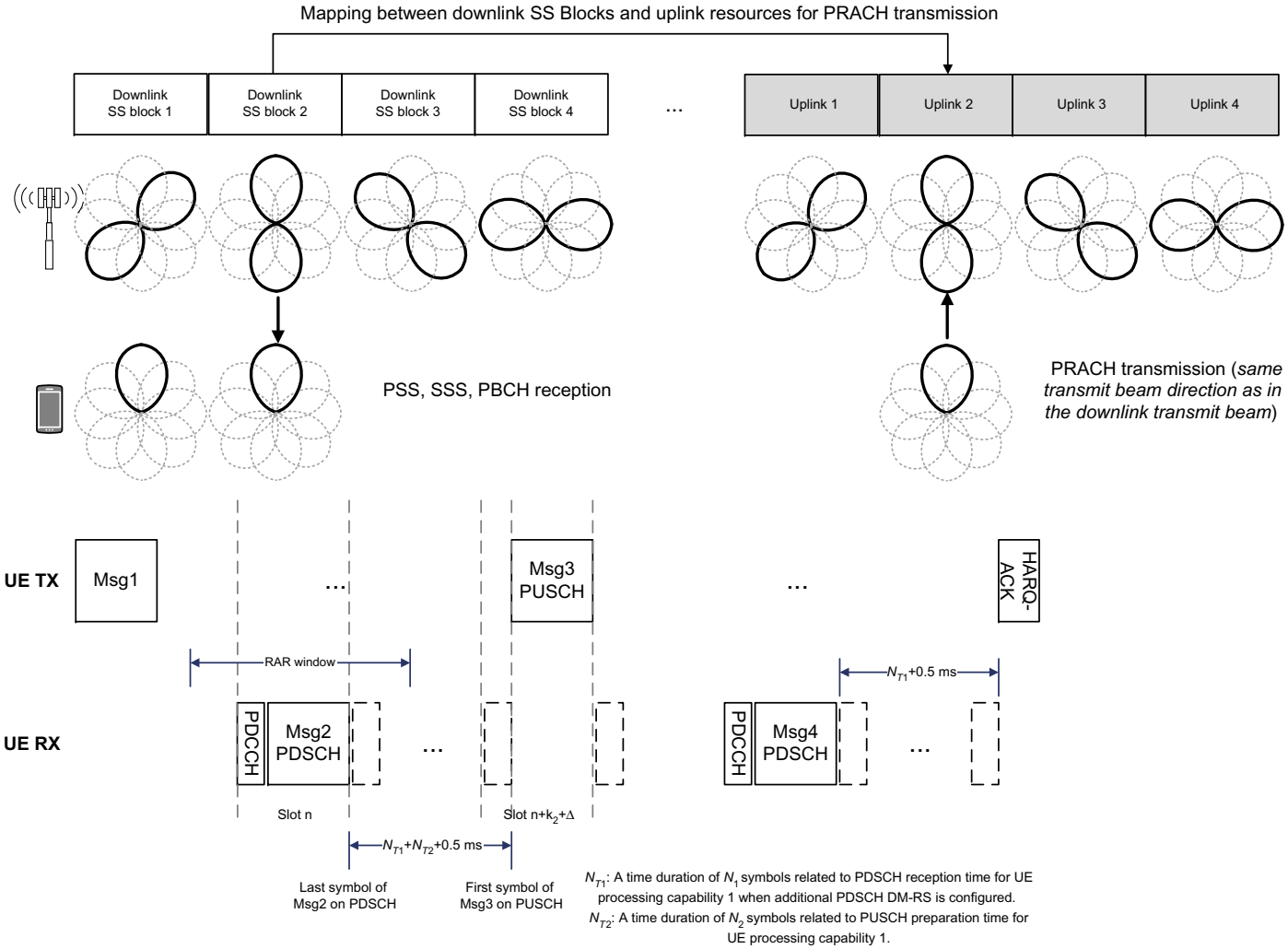


Figure 4.85
Random-access procedure [8,12].

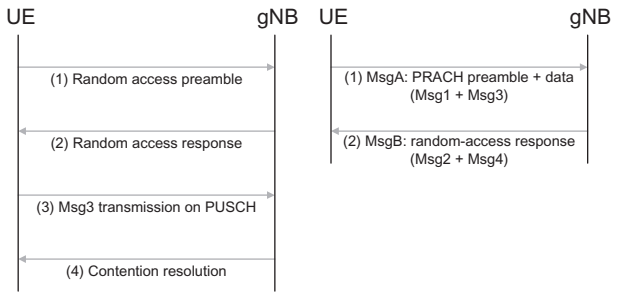


Figure 4.86
Comparison of two-step and four-step contention-based random-access procedures [11].

In NR Rel-15, the RACH procedure is triggered, when uplink data becomes available at the UE buffer and the UE is either in the RRC_IDLE/INACTIVE state, where the RACH procedure is triggered for state transition, or in the RRC_CONNECTED state, if the uplink is not synchronized, where the RACH procedure is used to reestablish uplink synchronization, or in the RRC_CONNECTED state, if the UE has no PUCCH resources available for SR or the SR procedure fails, where the RACH procedure serves as an SR. In addition, the RACH procedure is used for beam failure and recovery, on-demand SI request, or it can be explicitly triggered by the network with RRC for handover.

In NR Rel-15, the uplink data cannot be transmitted until the RACH procedure is successfully completed. It is observed that for small packet transmission, four-step RACH is not efficient in terms of latency and signaling overhead; thus the two-step RACH has been proposed to simplify the RACH procedure to achieve lower signaling overhead and latency. It is possible to allow Msg3 in four-step RACH procedure to carry data in order to reduce the latency and overhead. However, even in that case, the four-step RACH would still involve more signaling and latency relative to that of the two-step approach.

In two-step RACH, the MsgA may consist of two parts, that is, PRACH preamble and PUSCH which are time-division multiplexed. The PRACH preamble is used for UE detection, allowing the network to prepare for the reception of the corresponding PUSCH message. In NR Rel-15, up to 64 preamble signatures are mapped to one PRACH occasion. The preambles are orthogonal, or quasi-orthogonal, allowing the network to receive multiple preambles (from different UEs) in the same PRACH occasion. If all the preambles are mapped to a PUSCH in the same time-frequency occasion and more than one preamble is detected, the PUSCH transmissions of the detected preambles overlap in time and frequency, increasing the probability of PUSCH decoding failure. Alternatively, each preamble, or subset of preambles, can be mapped to a PUSCH in a unique time-frequency resource. This reduces the probability of PUSCH decoding failure due to collision but significantly increases the two-step RACH physical-layer overhead in the uplink. The MsgB in

two-step procedure comprises several fields including the detected unique ID for contention resolution, where the size of the detected ID depends on the use case; a timing advance field; an uplink-grant for scheduling the data packets after the RACH procedure; and small user-plane/control-plane packets for downlink communication. The presence and the size of each field depend on the use case; thus the total size of MsgB may vary.

4.2.4 Physical Uplink Shared Channel

The physical uplink shared channel is used to transmit the user traffic and control information in the uplink. It supports two transmission modes namely codebook-based and non-codebook-based multi-antenna transmission. For codebook-based transmission, the gNB provides the UE with a transmit precoding matrix indication in DCI. The UE uses the indicator to select the PUSCH transmit precoder from a set of codebooks. For non-codebook-based transmission, the UE determines its PUSCH precoder based on (wideband) SRS resource indication (SRI) field from DCI. A closed-loop DM-RS-based spatial multiplexing is supported for PUSCH with up to four transmission layers for SU-MIMO with CP-OFDM waveform. Uplink SU-MIMO uses one codeword. Support of DFT-S-OFDM in the uplink is optional, and when transform precoding is used, only a single MIMO transmission layer is supported.

As shown in Fig. 4.87, the uplink physical layer processing of transport channels consists of the following stages:

- Transport block CRC attachment, where TBSs larger than 3824 use 24-bit CRC and other TBSs utilize 16-bit CRC, followed by LDPC base graph selection
- Code block segmentation and code block CRC attachment, which always uses 24-bit CRC
- Channel coding which makes use of LDPC coding (base graph 1 or 2)
- Rate matching and code block concatenation followed by data and control multiplexing
- Scrambling and modulation where any of the modulation schemes may be used, that is, $\pi/2$-BPSK (with transform precoding only), QPSK, 16QAM, 64QAM, or 256QAM
- Layer mapping
- Transform precoding (enabled/disabled by configuration) and precoding
- Mapping to assigned resources and antenna ports

The UE transmits at least one symbol with DM-RS on each layer in which PUSCH is transmitted. The number of DM-RS symbols and resource element mapping is configured via RRC signaling. The PT-RS may be transmitted on additional symbols to assist the gNB receiver with phase tracking [11].

Following the above summary, let us discuss the physical layer processing of the UL-SCH in more detail. The CRC is calculated over the entire transport block that is constructed

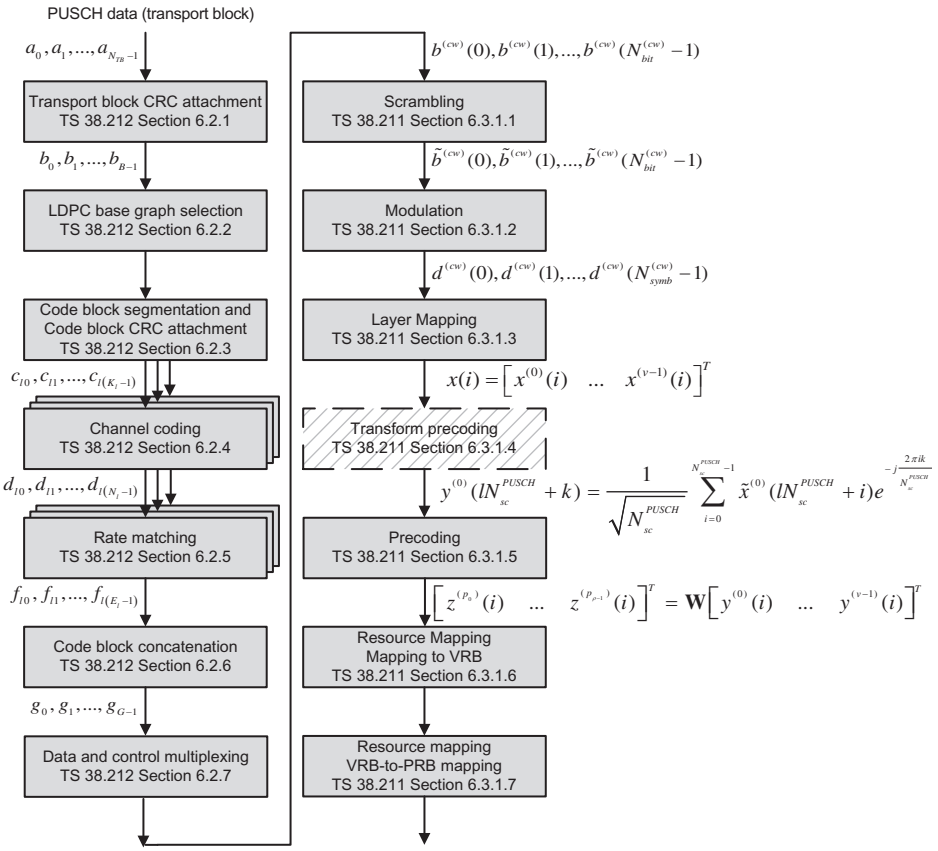


Figure 4.87
Physical processing of the uplink shared channel [30].

from the MAC PDU(s). We denote the TB bits by $a_0, a_1, \ldots a_{N_{PUSCH}-1}$ and the parity bits by $p_0, p_1, \ldots, p_{N_{CRC}-1}$, where N_{PUSCH} is the payload size and N_{CRC} is the number of parity bits. The number of parity bits depends on the PUSCH payload size. If $N_{PUSCH} > 3824$, $N_{CRC} = 24$ CRC bits are computed and attached to the TB using the generator polynomial $g_{CRC24A}(D) = D^{24} + D^{23} + D^{18} + D^{17} + D^{14} + D^{11} + D^{10} + D^7 + D^6 + D^5 + D^4 + D^3 + D + 1$; otherwise $N_{CRC} = 16$ CRC bits are calculated using the generator polynomial $g_{CRC16}(D) = D^{16} + D^{12} + D^5 + 1$. The code block bits after CRC attachment are denoted by $b_0, b_1, \ldots, b_{N_{PUSCH}+N_{CRC}-1}$ [7].

For initial transmission of a TB with coding rate R, determined by the MCS index and the subsequent retransmissions of the same TB, each code block of the TB is encoded with either LDPC base graph 1 or 2 depending on the values of N_{PUSCH} and R. If $N_{PUSCH} \leq 292$ or if $N_{PUSCH} \leq 3824$ and $R < 0.67$ or if $R < 0.25$, LDPC base graph 2 is used; otherwise, LDPC base graph 1 is used [7].

The input bit sequence to the code block segmentation is denoted by $b_0, b_1, \ldots, b_{N_{PUSCH}+N_{CRC}-1}$. If $B = N_{PUSCH} + N_{CRC}$ is larger than the maximum code block size K_{cb}, the bit sequence is segmented, and a 24-bit CRC is attached to each (segmented) code block. The value of K_{cb} for LDPC base graph 1 is $K_{cb} = 8448$ and for LDPC base graph 2, $K_{cb} = 3840$. The number of segmented code blocks is determined by $C = \lceil B/(K_{cb} - N_{CRC}) \rceil$. The output bits from code block segmentation are denoted by $c_{l0}, c_{l1}, \ldots, c_{l(K_l-1)}$ where $0 \leq l \leq C$ is the code block number, and $K_l = K$ is the number of bits in the lth code block.

The code blocks are then delivered to the channel coding unit where each code block is individually encoded with the LDPC encoder. The encoded bits are denoted by $d_{l0}, d_{l1}, \ldots, d_{l(N_l-1)}$, in which $N_l = 66Z_c$ for LDPC base graph 1 and $N_l = 50Z_c$ for LDPC base graph 2, where the value of the lifting factor Z_c is given in Table 4.7.

The encoded bits for each code block are processed through the rate matching function. The total number of code blocks is denoted by C, and each code block is individually rate matched by setting $I_{LBRM} = 1$, if RRC parameter *rateMatching* is set to *limitedBufferRM*; otherwise, by setting $I_{LBRM} = 0$. After the rate matching stage, the bits are denoted by $f_{l0}, f_{l1}, \ldots, f_{l(E_l-1)}$, where E_r is the number of rate matched bits in the lth code block.

The input bit sequence to the code block concatenation module are the sequences $\{f_{lk} | l = 0, 1, \ldots, C-1;\ \ k = 0, 1, \ldots, E_l - 1\}$ where E_l is the number of rate-matched bits in the lth code block. The output bit sequence from the code block concatenation function is the sequence $g_0, g_1, \ldots, g_{G-1}$ where G is the total number of coded bits for transmission. The code block concatenation function sequentially concatenate different rate-matched code blocks [7].

The NR supports UCI multiplexing on PUSCH when the UCI and PUSCH transmissions coincide in time, either due to transmission of an uplink TB or due to triggering of aperiodic CSI transmission without an uplink TB. The UCI carrying HARQ-ACK feedback with 1 or 2 bits is multiplexed by puncturing PUSCH. In all other cases, the UCI is multiplexed by rate matching PUSCH. In case of SUL, the UE is configured with two uplink carriers and one downlink carrier in the same cell. The uplink transmissions on the uplink carriers are controlled by the network to avoid overlapping PUSCH/PUCCH transmission in time-domain. Overlapping transmissions on PUSCH are avoided through scheduling, while overlapping transmissions on PUCCH are avoided through configuration since PUCCH can only be configured for one of the two uplink carriers in the cell. The initial access is supported on each uplink carrier. The mapping of the UCI to PUSCH resources is such that more (operationally) important bits (HARQ-ACK) are mapped to the first OFDM symbol after the first DM-RS, and less (operationally) important bits (CSI reports) are mapped to the subsequent symbols. Unlike the data part, which relies on rate adaptation to overcome the effects of radio propagation, the L1/L2 control signaling part cannot be rate-adapted. Power control may theoretically be used, but it would imply fast power variations in the time

domain, which negatively impact the RF properties. Therefore, the transmission power is maintained constant over the PUSCH duration and the amount of resource elements allocated to L1/L2 control signaling is changed by changing the code rate of the control signaling. In addition to a semi-static value controlling the amount of PUSCH resources used for UCI, it is also possible to signal this information as part of a DCI.

The bit sequence $b^{(q)}(0), b^{(q)}(1), \ldots, b^{(q)}(N_{bit}^{(q)} - 1)$ from the output of code block concatenation and multiplexing, where $N_{bit}^{(q)}$ is the number of bits in codeword q transmitted on the physical shared channel, are scrambled prior to modulation, resulting in a block of scrambled bits $\tilde{b}^{(q)}(0), \tilde{b}^{(q)}(1), \ldots, \tilde{b}^{(q)}(N_{bit}^{(q)} - 1)$ such that the UL-SCH bits (except the UCI place-holder bits) are scrambled with pseudo-random sequence $c(n)$ that is initialized with the 16-bit RNTI bits as follows $\tilde{b}^{(q)}(i) = (b^{(q)}(i) + c^{(q)}(i)) \bmod 2$, where $c_{init} = n_{RNTI}2^{15} + n_{ID}$ and $n_{ID} \in \{0, 1, \ldots, 1023\}$ is set equal to the RRC parameter *dataScramblingIdentityPUSCH*; otherwise, $n_{ID} = N_{ID}^{cell}$. The parameter n_{RNTI} corresponds to the RNTI associated with the PUSCH transmission, if the input bits corresponding to UCI place-holder bits are set to one or the previous scrambled bit, depending on the value of the bits [6].

The block of scrambled bits $\tilde{b}^{(q)}(0), \tilde{b}^{(q)}(1), \ldots, \tilde{b}^{(q)}(N_{bit}^{(q)} - 1)$ are modulated using $\pi/2$-BPSK (with transform precoding only), QPSK, 16QAM, 64QAM, or 256QAM modulation schemes, resulting in a block of complex-valued modulation symbols $d^{(q)}(0), d^{(q)}(1), \ldots, d^{(q)}(N_{symb}^{(q)} - 1)$.

The complex-valued modulation symbols can be mapped to a maximum of four layers. More specifically, the complex-valued modulation symbols are mapped to $x(i) = [x^{(0)}(i) \ldots x^{(v-1)}(i)]^T$ layer, $i = 0, 1, \ldots, N_{layer}^{symb}$ where v is the number of layers and N_{layer}^{symb} is the number of modulation symbols per layer. If transform precoding is not enabled; for uplink CP-OFDM, $y^{(\lambda)}(i) = x^{(\lambda)}(i)$ for each layer $\lambda = 0, 1, \ldots, v - 1$. However, for DFT-S-OFDM uplink, $v = 1$, and $\tilde{x}^{(0)}(i)$ depends on the configuration of phase-tracking reference signals. If phase-tracking reference signals are not configured, the block of complex-valued symbols $x^{(0)}(0), \ldots, x^{(0)}(N_{layer}^{symb} - 1)$ for the single-layer $v = 1$ are divided into $N_{layer}^{symb}/N_{sc}^{PUSCH}$ sets, each corresponding to one OFDM symbol and $\tilde{x}^{(0)}(i) = x^{(0)}(i)$. In case phase-tracking reference signals are configured, the block of complex-valued symbols $x^{(0)}(0), \ldots, x^{(0)}(N_{layer}^{symb} - 1)$ are divided into a number of groups, where each group corresponds to one OFDM symbol. The lth group contains $N_{sc}^{PUSCH} - \varepsilon_l N_{samp}^{group} N_{group}^{PT\text{-}RS}$ subcarriers and is mapped to the complex-valued symbols $\tilde{x}^{(0)}(lN_{sc}^{PUSCH} + i')$ corresponding to the lth OFDM symbol prior to transform precoding, wherein $i' = \{0, 1, \ldots, N_{sc}^{PUSCH} - 1\}$ and $i' \neq m$. The index m of PT-RS samples in the lth group, the number of samples per PT-RS group N_{samp}^{group}, and the number of PT-RS groups $N_{group}^{PT\text{-}RS}$ are defined in [6]. The quantity $\varepsilon_l = 1$, when the lth OFDM symbol contains one or more PT-RS samples, otherwise $\varepsilon_l = 0$.

The transform precoding is then performed, resulting in a block of complex-valued symbols $y^{(0)}(0),\ldots,y^{(0)}(N_{layer}^{symb}-1)$ as follows [6]:

$$y^{(0)}\left(lN_{sc}^{PUSCH}+k\right)=\frac{1}{\sqrt{N_{sc}^{PUSCH}}}\sum_{i=0}^{N_{sc}^{PUSCH}-1}\tilde{x}^{(0)}\left(lN_{sc}^{PUSCH}+i\right)e^{-j\left(2\pi ik/N_{sc}^{PUSCH}\right)};$$

$$k=0,\ldots,N_{sc}^{PUSCH}-1;\quad l=0,\ldots,\frac{N_{layer}^{symb}}{N_{sc}^{PUSCH}}-1$$

The parameter $N_{sc}^{PUSCH}=N_{RB}^{PUSCH}N_{sc}^{RB}$ where N_{RB}^{PUSCH} represents the bandwidth of the PUSCH in terms of resource blocks which must satisfy $N_{sc}^{PUSCH}=2^{\alpha_2}3^{\alpha_3}5^{\alpha_5}$ where $\alpha_2,\alpha_3,\alpha_5$ are non-negative integers. The DFT precoding is used to reduce the cubic metric of the uplink signal, thereby enabling higher power-amplifier efficiency. From implementation point of view, it is better to constrain the DFT size to a power of 2. However, such a constraint would limit the scheduler flexibility in terms of the amount of resources that can be assigned for an uplink transmission. In NR, the DFT precoding size, and thus the size of the resource allocation, is limited to products of the integers 2, 3, and 5 such that the DFT can be implemented as a combination of relatively less complex radix-2, radix-3, and radix-5 FFT processing.

The block of vectors $[y^{(0)}(i)\ldots y^{(v-1)}(i)]^T$, $i=0,1,\ldots,N_{layer}^{symb}-1$ corresponding to layers are precoded as follows:

$$\left[z^{(p_0)}(i)\ldots z^{(p_{\rho-1})}(i)\right]^T=\mathbf{W}\left[y^{(0)}(i)\ldots y^{(v-1)}(i)\right]^T$$

where $i=0,1,\ldots,N_{ap}^{symb}-1$, $N_{ap}^{symb}=M_{layer}^{symb}$, and p_i denotes the antenna port. For non-codebook-based transmission, the precoding matrix $\mathbf{W}$ is an identity matrix. However, for codebook-based transmission, the precoding matrix $\mathbf{W}$ is a scaler equal to one for single-layer transmission on a single-antenna port; otherwise, depending on value of the transmitted precoding matrix indicator (TPMI) index obtained from the DCI scheduling the uplink transmission, it will be chosen from a set of predefined matrices [6]. As an example, Table 4.32 provides the entries of the precoding vectors for single-layer transmission using two antenna ports.

Table 4.32: Precoding matrix W for single-layer transmission using two antenna ports [6].

	TPMI Index 0	TPMI Index 1	TPMI Index 2	TPMI Index 3	TPMI Index 4	TPMI Index 5
W	$\frac{1}{\sqrt{2}}\begin{bmatrix}1\\0\end{bmatrix}$	$\frac{1}{\sqrt{2}}\begin{bmatrix}0\\1\end{bmatrix}$	$\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$	$\frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}$	$\frac{1}{\sqrt{2}}\begin{bmatrix}1\\j\end{bmatrix}$	$\frac{1}{\sqrt{2}}\begin{bmatrix}1\\-j\end{bmatrix}$

For each antenna port used for transmission of PUSCH, the block of complex-valued symbols $z^{(p)}(0), \ldots, z^{(p)}(M_{ap}^{symb} - 1)$ are scaled by a factor of β_{PUSCH} to adjust the transmit power and sequentially mapped, starting with $z^{(p)}(0)$, to virtual resource elements (k', l) in the virtual resource blocks assigned for PUSCH transmission. The physical resource blocks corresponding to the latter virtual resources must not be used for transmission of DM-RS, PT-RS, or DM-RS intended for other co-scheduled UEs. The mapping to virtual resource elements (k', l) is in increasing order of frequency index k' over the assigned virtual resource blocks, where $k' = 0$ is the first subcarrier in the lowest numbered virtual resource block followed by time index l. The virtual resource blocks are then mapped to physical resource blocks in a non-interleaved manner. For non-interleaved VRB-to-PRB mapping, virtual resource block n is mapped to physical resource block n.

While dynamic scheduling is the basic mode of operation in NR, the resources for uplink data transmission or downlink data reception can be configured in advance for the UE. Once the uplink data are available at UE's buffer, it can immediately start uplink transmission without going through the SR and grant cycle, thus reducing the latency. In other words, the NR PUSCH transmissions can be dynamically scheduled by an uplink grant provided by a DCI, or the transmission can correspond to a configured grant Type 1 or Type 2. As shown in Fig. 4.88, the configured grant Type 1 PUSCH transmission is semi-statically configured to operate upon the reception of RRC parameter *configuredGrantConfig* including *rrc-ConfiguredUplinkGrant* without the detection of an uplink grant in a DCI. The configured grant Type 2 PUSCH transmission is semi-persistently scheduled by an uplink grant in a valid activation DCI after the reception of RRC parameter *configurdGrantConfig* that does not include *rrc-ConfiguredUplinkGrant*. The UE transmits PUSCH upon detection of a PDCCH with DCI format 0_0 or 0_1, and it is not expected to be scheduled to transmit

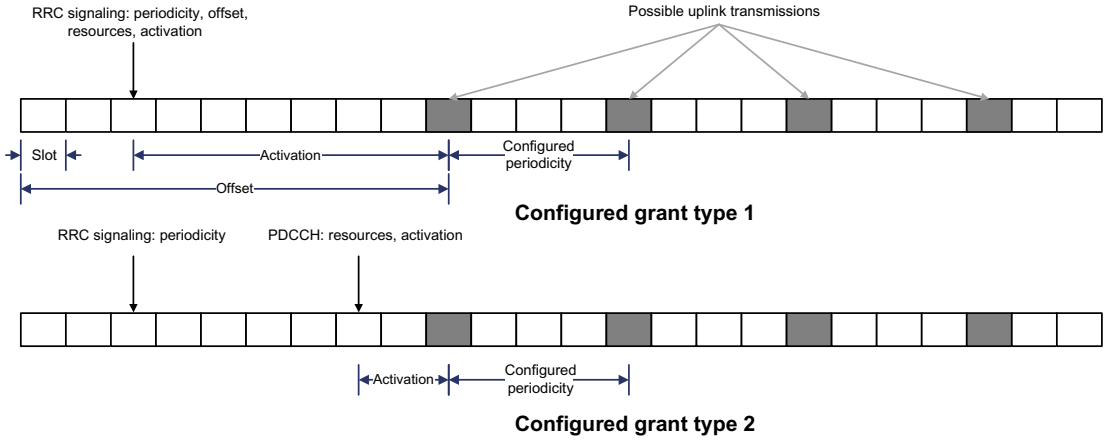


Figure 4.88
Illustration of uplink transmission with configured grants Type 1 and Type 2 [9].

another PUSCH by DCI format 0_0 or 0_1 scrambled by C-RNTI or MCS-C-RNTI for a given HARQ process until the end of the expected transmission of the last PUSCH for that HARQ process [9].

When the UE is scheduled to transmit a TB without a CSI report, or it is scheduled to transmit a TB with CSI report(s) on PUSCH by a DCI, the *time domain resource assignment* field value m of the DCI provides a row index $m+1$ to an allocated table. The indexed row defines the slot offset K_2, the start and length indicator *SLIV*, or directly the start symbol S and the allocation length L, and the PUSCH mapping type to be applied. Alternatively, when a UE is scheduled to transmit a CSI report(s) on PUSCH without a TB, the *time domain resource assignment* field value m of the DCI provides a row index $m+1$ to an allocated table which is defined by RRC parameter *pusch-TimeDomainAllocationList* in *pusch-Config*. The indexed row defines the start and length indicator *SLIV*, and the PUSCH mapping type. The parameter K_2 value is determined as $K_2 = \max_j Y_j(m+1)$, where $Y_j, j = 0, \ldots, N_{repetition} - 1$ are the corresponding list entries of the RRC parameter *reportSlotOffsetList* in *CSI-ReportConfig* for the $N_{repetition}$ triggered CSI Reporting Settings and $Y_j(m)$ is the mth entry of Y_j [9]. The slot where the UE is expected to transmit PUSCH is determined by $\lfloor (2^{\mu PUSCH}/2^{\mu PDCCH})n \rfloor + K^2$ where n is the slot with the scheduling DCI, K_2 is based on PUSCH numerology, and μ_{PUSCH} and μ_{PDCCH} are the subcarrier spacing for PUSCH and PDCCH, respectively. The starting symbol S relative to the start of the slot, and the number of consecutive symbols L counting from symbol S allocated for PUSCH are determined from the start and length indicator *SLIV* of the indexed row such that if $(L-1) \leq 7$ then $SLIV = 14(L-1) + S$; otherwise $SLIV = 14(15-L) + (13-S)$ where $0 < L \leq 14 - S$. The PUSCH mapping type is set to Type A or Type B as given by the indexed row [9]. We have mentioned before that the time-domain reference point of the first PUSCH DM-RS symbol depends on the mapping type, where for PUSCH mapping type A, l is defined relative to the start of the slot, if frequency hopping is disabled and relative to the start of each hop in case frequency hopping is enabled, and the offset l_0 is given by the higher layer parameter *dmrs-TypeA-Position*. For PUSCH mapping type B, time-domain index l is defined relative to the start of the scheduled PUSCH resources, if frequency hopping is disabled and relative to the start of each hop when frequency hopping is enabled and the offset $l_0 = 0$ [6].

The UE determines the resource block assignment in frequency domain using the resource allocation field in the detected PDCCH DCI except for Msg3 PUSCH initial transmission. Two uplink resource allocation types are supported. The uplink resource allocation Type 0 is supported for PUSCH only when transform precoding is disabled, whereas the uplink resource allocation Type 1 is supported for PUSCH regardless of whether transform precoding is disabled. If the scheduling DCI is configured to indicate the uplink resource allocation type as part of the *Frequency domain resource* assignment field by setting RRC parameter *resourceAllocation* in *pusch-Config* to "dynamicswitch", the UE must use uplink

resource allocation Type 0 or Type 1; otherwise, it uses the uplink frequency resource allocation type as defined by the RRC parameter *resourceAllocation* [9]. When the UE is scheduled with DCI format 0_0, the uplink resource allocation Type 1 is used. If a bandwidth part indicator field is not configured in the scheduling DCI, the RB indexing for uplink Type 0 and Type 1 resource allocation is determined within the UE's active bandwidth part. However, if a bandwidth part indicator field is configured in the scheduling DCI, the RB indexing for uplink Type 0 and Type 1 resource allocation is determined within the UE's bandwidth part indicated by bandwidth part indicator field value in the DCI. Upon detection of PDCCH, the UE first determines the uplink bandwidth part and then the resource allocation within the bandwidth part where RB numbering starts from the lowest RB in the determined uplink bandwidth part [9].

4.2.5 Uplink MIMO Schemes

The NR supports multi-antenna precoding with up to four layers for PUSCH transmission. However, when uplink DFT-based transform precoding is enabled, only single-layer (rank 1) transmission is supported. The UE can be configured in either codebook-based or non-codebook-based modes for PUSCH transmission. The selection between these two modes depends on the extent to which the uplink channel conditions can be estimated by the UE based on downlink measurements. PUSCH DM-RSs are precoded in the same way that PUSCH data subcarriers are precoded to allow coherent demodulation, making uplink precoding transparent to the gNB receiver. When codebook-based precoding is used in the uplink, the scheduling grant includes information about the precoder in the same way that the UE provides the network with PMI to assist downlink multi-antenna precoding. However, in contrast to the downlink, where the network may or may not use the precoder matrix indicated by the PMI, in the uplink direction, the UE is expected to use the precoder suggested by the network. In case of non-codebook-based transmission, the network can influence the selection of the uplink precoder. Another aspect that may impose constraints on the uplink multi-antenna transmission is to what extent one can assume coherence between different device antennas, that is, to what extent the relative phase between the signals transmitted on two antennas can be controlled. The phase coherence is necessary when antenna port−specific weight factors, including specific phase shifts, are applied to the signals transmitted on the different antenna ports. The NR specifications allow different UE capabilities concerning inter-antenna-port phase coherence, referred to as full coherence, partial coherence, and no coherence. In case of full coherence, it can be assumed that the device can control the relative phase between any of its antenna ports that are used for uplink transmission. In case of partial coherence, the device is capable of pairwise coherence, that is, the device can control the relative phase between antenna-port pairs. However, there is no guarantee that coherence can be achieved. In case of no coherence, there is no guarantee of phase coherence between any pair of the device antenna ports [6,9,14].

In codebook-based uplink shared-channel transmission, the network selects the transmission rank and the corresponding precoding matrix and informs the device through uplink scheduling grant. At the UE side, the precoding matrix is applied to the scheduled PUSCH transmission and the indicated number of layers is mapped to the antenna ports. To select a suitable rank and a corresponding precoding matrix, the gNB needs estimates of the channels between the device antenna ports and the corresponding network receive antennas. To enable this, a UE configured for codebook-based PUSCH transmission would typically be configured for transmission of at least one multi-port SRS. Based on measurements on the configured SRS, the network can estimate the channel and determine a suitable rank and precoding matrix. The network cannot select an arbitrary precoder, rather for a given combination of the antenna ports and transmission rank, the network can select the precoding matrix from a limited set of available precoders [6,9,14].

When selecting the precoding matrix, the network needs to consider the device capability in terms of antenna-port phase coherence. If the UE does not support antenna-port phase coherence, only the first two precoding matrices can be used with rank-1 transmission. It must be noted that restricting the codebook selection to these two matrices is equivalent to selecting either the first or the second antenna port for transmission. In this type of antenna selection, phase coherence between the antenna ports is not required. The selection of the remaining precoding vectors would imply linear combination of the signals of different antenna ports, which requires phase coherency among the antenna ports. A fundamental difference between NR codebook-based PUSCH transmission and LTE uplink is that a device can be configured to transmit multiple SRS from multiple antenna ports. In multi-SRS transmission, the network feedback includes SRI, identifying one of the configured SRSs. The UE should then use the precoder identified in the scheduling grant and map the output of the precoder to the antenna ports corresponding to the SRSs indicated in the SRI. The device should then transmit the precoded signal using the same antenna configuration and mapping that was used for the SRS transmission (indicated by the SRI). The use of multiple SRSs for codebook-based PUSCH transmission assumes that the UE transmits multi-port SRSs over separate and relatively wide beams. These beams may correspond to different UE antenna panels with different directions, where each panel includes a set of antenna elements corresponding to the antenna ports of each multi-port SRS (see Fig. 4.89). The SRI received from the network determines which beam should be used for the transmission, while the precoder information (number of layers and precoder) determines how the transmission should be done within the selected beam. Codebook-based precoding is typically used when uplink/downlink reciprocity cannot be achieved and when uplink measurements are needed in order to determine a suitable uplink precoding [6,9,14].

In contrast to codebook-based precoding, which is based on network measurements and selection of uplink precoder, non-codebook-based precoding is based on device

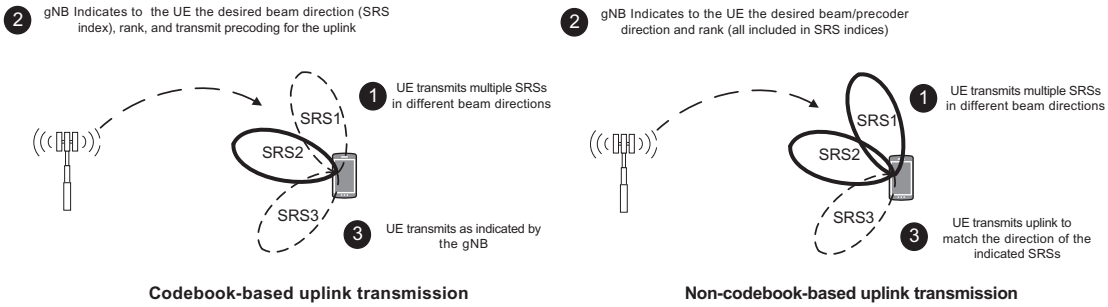


Figure 4.89
Codebook-based uplink transmission versus non-codebook-based uplink transmission [69].

measurements and precoder indications to the network. The concept of uplink non-codebook-based precoding is illustrated in Fig. 4.89. Based on downlink measurements conducted on configured CSI-RS resources, the UE selects a suitable uplink multi-layer precoder. Non-codebook-based precoding relies on channel reciprocity and assumes that the device can acquire accurate knowledge of the uplink channel based on downlink measurements. Note that there are no restrictions on the selection of precoder by the UE. Each column of a precoding matrix defines a digital beam for the corresponding layer. Therefore, selection of precoder for each layer can be perceived as selection of different beam directions, where each beam corresponds to one possible layer. It must be noted that the UE precoder selection is typically done based on downlink measurements, which may not be necessarily the best precoder from network point of view. As a result, the NR non-codebook-based precoding includes an additional step where the network can modify the device-selected precoder by removing some of the beams or equivalently some columns from the selected precoder [6,9,14].

As we mentioned earlier, codebook-based and non-codebook-based transmission modes are supported for PUSCH transmission. The UE will be configured with codebook-based transmission, when the RRC parameter *txConfig* in *pusch-Config* is set to "*codebook*", and it will be configured with non-codebook-based transmission, when the RRC parameter *txConfig* is set to "*nonCodebook*". If the RRC parameter *txConfig* is not provided, the PUSCH transmission will be based on a single-antenna port, triggered by DCI format 0_0. For codebook-based transmission, the UE determines the transmission precoder based on the information obtained from the SRI, TPMI, and the transmission rank, where the SRI, TPMI, and the transmission rank are given by the corresponding fields of the DCI. The TPMI is used to identify the preferred precoder over the SRS ports in the selected SRS resource by the SRI when single or multiple SRS resources are configured. Note that the indicated SRI in *n*th slot is associated with the most recent transmission of SRS resource identified by the SRI, where the SRS resource is prior to the PDCCH carrying the SRI. In codebook-based transmission mode, the UE determines its codebook subsets based on

TPMI and following reception of RRC parameter *codebookSubset* in *pusch-Config,* which may be configured with "*fullyAndPartialAndNonCoherent*", or "*partialAndNonCoherent*", or "*nonCoherent*" depending on the UE capability. The maximum transmission rank may be configured by the higher parameter *maxRank* in *pusch-Config.* Furthermore, for codebook-based transmissions, the UE may be configured with a single *SRS-ResourceSet*, and only one SRS resource can be indicated based on the SRI in the SRS resource set. The maximum number of configured SRS resources for codebook-based transmission is 2. If aperiodic SRS is configured for a UE, the SRS request field in DCI triggers the transmission of aperiodic SRS resources [9].

The codebook subset restriction concept was introduced in LTE [15]. It helps avoid CSI reporting for the undesired (spatial) directions. The LTE codebook subset restriction includes RI and PMI restriction, which provides sufficient flexibility to control PMI calculation and transmission from the UE. The content of CSI in NR is very similar to LTE, in the sense that NR supports CSI components such as RI and PMI.

Since the number of possible RI values is small, bitmap with one-to-one correspondence between each bit in bitmap, and RI value can be specified for the purpose of RI restriction. At the same time, the number of possible PMI values especially for larger number of antenna ports is very large to support one-to-one correspondence between PMIs and bits in the bitmap. Therefore, a solution with reduced signaling overhead should be considered. More specifically, similar to LTE, a DFT beam restriction is introduced, so that the PMI can be considered as restricted, if at least one beam is restricted by the corresponding DFT beam restriction bitmap. For co-phasing of the polarization, the bitmap should not be used as it does not affect the beamforming direction.

The NR Type-I single-panel codebook structure is similar to that of LTE FD-MIMO codebooks, except rank 3/4 codebooks for 16, 24, and 32 antenna ports at the gNB. Let us consider codebook subset restriction for less than 16 antenna ports at the gNB. In this case, the beamforming vector for PMIs of all ranks is represented by 2D DFT beam denoted as $\mathbf{b}_i$, which is represented as Kronecker product of two 1D DFT vectors $\mathbf{b}_i = \mathbf{u}_n \otimes \mathbf{q}_l$ wherein [66]

$$\mathbf{u}_n = \frac{1}{\sqrt{N_1}}\left[1 \quad e^{2\pi j\left(1/N_1O_1\right)n} \quad \cdots \quad e^{2\pi j\left((N_1-1)/N_1O_1\right)n}\right]^T$$

$$\mathbf{q}_l = \frac{1}{\sqrt{N_2}}\left[1 \quad e^{2\pi j\left(1/N_2O_2\right)l} \quad \cdots \quad e^{2\pi j\left((N_2-1)/N_2O_2\right)l}\right]^T$$

In this case, in order to restrict transmission in specific direction, a bitmap with size $N_1N_2O_1O_2$ can be specified, where each bit a_i corresponds to DFT beam $\mathbf{b}_i$. If at least one layer of the PMI consists of $\mathbf{b}_i$, the PMI is considered to be restricted and cannot be

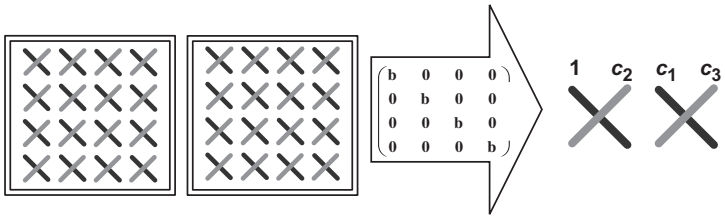


Figure 4.90
Illustration of precoding for multi-panel antenna array [66].

reported by the UE. Rank 3/4 PMIs for 16, 24, and 32 antenna ports have different structures compared to Type-I single-panel PMI. Therefore, it may not be possible to use the same approach for all ranks. The multi-panel codebook is constructed by DFT-based beamforming per each panel and co-phasing of polarization and panels, where the same DFT beam is applied for all panels and polarizations. An example is shown in Fig. 4.90, where precoder $\mathbf{p}$ for rank-1 multi-panel codebook and two-panel antenna can be computed as follows, in which c_1, c_2, c_3 coefficients are independently reported in accordance to mode 2 of multi-panel codebooks [66].

$$\mathbf{p} = \begin{pmatrix} \mathbf{b} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{b} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{b} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{b} \end{pmatrix} \begin{pmatrix} 1 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ c_1\mathbf{b} \\ c_2\mathbf{b} \\ c_3\mathbf{b} \end{pmatrix}$$

In the preceding example, it is assumed that the antenna port indexing is performed in such way that $[\mathbf{b} \quad c_1\mathbf{b}]$ corresponds to beamforming vector of the first polarization and $[\mathbf{b} \quad c_3/c_2\mathbf{b}]$ to beamforming vector of the second polarization.

The direction of the transmission in the above PMI structure is determined by the DFT beam denoted by vector $\mathbf{b}$ and co-phasing coefficients c_2 and c_3/c_2. Therefore, codebook subset restriction for multi-panel codebook PMI restriction consider all possible combinations of DFT beams and co-phasing coefficients, which are determining the direction of the transmission. The resulting size of bitmap in that case equals to $N_1N_2O_1O_2 4^{(Ng-1)}$ [66].

The Type-II CSI was designed to enhance the performance of MU-MIMO transmission. The accuracy of spatial channel feedback in case of Type-II CSI allows interference suppression improvement through use of advanced precoding schemes such MMSE precoding. An accurate knowledge of the channel increases suppression capabilities of intra-cell interference. The beamforming vector in Type-II codebook is represented as linear combination of 2, 3, or 4 DFT beams as $\mathbf{w}_{rl} = \sum_{i=0}^{L-1} p_{rli}^{(WB)} p_{rli}^{(SB)} c_{rli} \mathbf{v}_{k_i}$, where $p_{rli}^{(WB)}$ denotes the wideband beam amplitude scaling factor; $p_{rli}^{(SB)}$ is the subband beam amplitude scaling factor, and c_{rli} is the beam combining coefficient (phase) for beam i, polarization r, and layer l [66].

For non-codebook-based uplink transmission, PUSCH transmission can be scheduled by DCI format 0_0, DCI format 0_1 or semi-statically configured. The UE can determine its PUSCH precoder and transmission rank based on the SRI when multiple SRS resources are configured, where the SRI is given by the SRI in DCI or the SRI is given by RRC parameter *srs-ResourceIndicator*. The UE must use one or multiple SRS resources for SRS transmission. In an SRS resource set, the maximum number of SRS resources which can be configured for the UE for simultaneous transmission on the same symbol and the maximum number of SRS resources depend on the UE capability. It must be noted that only one SRS port for each SRS resource can be configured and only one SRS resource set can be configured with higher layer parameter *usage* in *SRS-ResourceSet* set to "*nonCodebook*". The maximum number of SRS resources that can be configured for non-codebook-based uplink transmission is 4. The indicated SRI in the *n*th slot is associated with the most recent transmission of SRS resource(s) identified by the SRI, where the SRS transmission is prior to the PDCCH carrying the SRI [9].

For non-codebook-based uplink transmission, the UE can calculate the precoder used for the transmission of SRS based on measurement of an associated NZP CSI-RS resource. A UE can be configured with only one NZP CSI-RS resource for the SRS resource set with higher layer parameter *usage* in *SRS-ResourceSet* set to "*nonCodebook*". If aperiodic SRS resource set is configured, the associated NZP-CSI-RS is indicated via SRS request field in DCI formats 0_1 and 1_1. A UE is not expected to update the SRS precoding information, if the gap between the last symbol of the reception of the aperiodic NZP-CSI-RS resource and the first symbol of the aperiodic SRS transmission is less than 42 OFDM symbols. The CSI-RS is located in the same slot as the SRS request field. If the UE is configured with aperiodic SRS associated with aperiodic NZP CSI-RS resource, none of the TCI states configured on the scheduled component carrier are configured with "QCL-TypeD". The UE performs one-to-one mapping between the indicated SRI(s) and the indicated DM-RS port (s) and their corresponding PUSCH layers provided by DCI format 0_1 or by *configuredGrantConfig*. The UE transmits PUSCH using the same antenna ports as the SRS port(s) in the SRS resource(s) indicated by SRI(s) given by DCI format 0_1 or by *configuredGrantConfig*. For non-codebook-based uplink transmission, the UE can be scheduled with DCI format 0_1 when at least one SRS resource is configured in *SRS-ResourceSet* with *usage* set to "*nonCodebook*" [9].

4.2.6 Link Adaptation and Power Control

Power control is a mechanism where the transmit power of the downlink or uplink control or traffic channels are adjusted at the gNB or at UEs, based on instructions from the serving base station such that with minimal impact on the reliability of the downlink/uplink transmissions and throughput, the inter-user/inter-cell interference among users and base stations

are reduced. Therefore, power control can be considered as a link adaptation mechanism that is utilized for interference mitigation in cellular systems.

While increasing the transmit power over a communication link has certain advantages such as higher SNR at the receiver, which reduces the BER and allows higher data rate and results in greater spectral efficiency as well as more protection against signal attenuation over fading channels, a higher transmit power; however, has several drawbacks including increased power consumption of the transmitting device, reducing the UE battery life, and increased interference to other users in the same or adjacent frequency bands. The following sections describe the power control algorithms that are incorporated in NR.

The NR provides uplink power control mechanisms to compensate the effects of path loss, shadowing, fast fading, and implementation loss. The uplink power control is further used to mitigate inter-cell and intra-cell interference, thereby enhancing the overall throughput and reducing the effective UE power consumption. The uplink power control includes open-loop and closed-loop power control. The base station transmits necessary power control information through transmission of power control messages. The parameters of the power control algorithm are optimized on a system-wide basis by the gNB and are broadcast periodically or trigged by events. The UE provides the necessary information through higher layer control messages to the serving gNB in order to enable uplink power control. The gNB can exchange necessary information with neighboring base stations through backhaul to support uplink power control to facilitate the handover process.

The power control scheme may not be effective in high mobility scenarios for compensating the effects of a fast fading channel due to variation of the channel impulse response. As a result, the power control is used to mitigate the distance-dependent path loss, shadowing, and implementation loss. The uplink power control takes into consideration the MIMO transmission mode and whether a single user or multiple users are supported on the same resource at the same time. The open-loop power control compensates the channel variations and implementation loss without requiring frequent interactions with the serving gNB. The UE can determine the transmit power based on the transmission parameters sent by the gNB, uplink channel quality, downlink CSI, or the interference knowledge obtained from downlink transmissions. The open-loop power control provides a coarse initial transmit power setting for the device before establishing connection with the base station. It is believed that rate control is more efficient than power control under certain conditions. Rate control in principle implies that the power amplifier is always transmitting at full power and therefore is efficiently utilized. On the other hand, power control often results in inefficient utilization of the power amplifier because the transmission power is often less than its maximum. In practice, the radio-link data rate is controlled by adjusting the modulation scheme and/or the channel coding rate. In good channel conditions, the value of E_b/N_0 at the receiver is high, and the main limitation of the data rate is the bandwidth of the radio link.

In such conditions, use of higher order modulation, for example, 16QAM 64QAM, or 256QAM together with a high coding rate, is more appropriate for link adaptation. Similarly, in the case of poor channel conditions, the use of QPSK and low-rate coding is preferred. Link adaptation by means of rate control is referred to as adaptive modulation and coding.

A power control mechanism takes into consideration the serving gNB target link SINR and/or interference level to other cells/sectors for mitigating inter-cell interference. In order to achieve the target SINR, the serving gNB path loss can be fully or partially compensated based on a trade-off between overall system throughput and cell-edge performance. The UE transmit power is adjusted in order to ensure the level of interference is less than the permissible interference level. The closed-loop power control, on the other hand, compensates channel variations through periodic power-control commands from the serving gNB. The base station measures the uplink CSI and interference level using uplink data and/or control channel transmissions and sends power control commands to the devices. Upon receiving the power control command from the gNB, the UE adjusts its uplink transmit power. The closed-loop power control is active during data and control channel transmissions. A UE is expected to maintain the transmit power density (i.e., total transmit power normalized by transmission bandwidth) for each data and control channel below a certain level that is determined by the maximum permissible power level for the UE, emission mask, and other regulatory constraints. In other words, when the number of active logical resource units assigned to a particular user is reduced, the total transmitted power must be reduced proportionally by the UE in the absence of any additional change of power control parameters.

When the number of resource blocks is increased, the total transmitted power must be proportionally increased such that the transmitted power level does not exceed the permissible power levels specified by 3GPP and the regulatory specifications [1–3]. For interference level control, the information about the current interference level of each gNB may be shared among the base stations via backhaul.

The (uplink) TPC in mobile communication systems is meant to balance the transmitted energy per bit in order to maintain the link quality corresponding to the minimum QoS requirements, to minimize interference to other users in the system, and to minimize the power consumption of the device. In achieving these goals, the power control has to adapt to the characteristics of the propagation channel, including path loss, shadowing, and fast fading, as well as overcoming interference from other users both within the same cell and in neighboring cells. The NR uplink power control is similar to LTE and is based on a combination of open-loop power control, including support for fractional path loss compensation, where the device estimates the uplink path loss based on downlink measurements and sets the transmit power accordingly, and closed-loop power control based on explicit power-control commands provided by the network. In practice, these power-control

commands are determined based on prior measurements of the received uplink power. The main difference with the LTE control is the possibility of beam-based power control.

Uplink power control determines the power level for PUSCH, PUCCH, SRS, and PRACH transmissions. The *i*th PUSCH/PUCCH/SRS/PRACH transmission occasion is defined by slot index n_{slot} within a frame with system frame number *SFN*, the first symbol *S* within the slot, and the number of consecutive symbols *L*. For a PUSCH transmission on active uplink BWP k of carrier l of serving cell m and parameter set configuration with index j and PUSCH power control adjustment state with index u, a UE first calculates a linear value $\hat{P}^{klm}_{PUSCH}(i,j,q_d,u)$ of the transmit power $P^{klm}_{PUSCH}(i,j,q_d,u)$. If PUSCH transmission is scheduled by DCI format 0_1 and when *txConfig* in *PUSCH-Config* is set to "*codebook*", the UE scales the linear value by the ratio of the number of antenna ports with a non-zero PUSCH transmission power to the maximum number of SRS ports supported by the UE in one SRS resource. The UE divides the power equally among the antenna ports on which it transmits PUSCH with non-zero power. The PUSCH transmit power is calculated as follows [8]:

$$P^{klm}_{PUSCH}(i,j,q_d,u) = \min\Big\{ P_{CMAX_{lm}}(i)P_{o\_PUSCH_{klm}}(j) + 10\log_{10}\left[2^{\mu}\mathrm{N}^{PUSCH}_{RB_{klm}}(i)\right] + \alpha_{klm}(j)PL_{klm}(q_d) + \Delta TF_{klm}(i) + f_{klm}(i,u)\Big\} \text{ (dBm)}$$

where

- (i,j,q_d,u) denote transmission occasion, parameter set configuration index, reference signal index for the active downlink BWP, and PUSCH power control adjustment state index, respectively.
- $P_{CMAX_{lm}}(i)$ denotes the maximum permissible UE transmit power for carrier l, serving cell m, and in PUSCH transmission occasion i.
- $P_{0\_PUSCH_{klm}}(j) = P_{0\_NOMINAL\_PUSCH_{lm}}(j) + P_{0\_UE\_PUSCH_{klm}}(j)$; $j \in \{0,1,\ldots,J-1\}$ is a parameter determined by RRC parameters *preambleReceivedTargetPower*, *msg3-DeltaPreamble*, *ConfiguredGrantConfig*, *p0-NominalWithGrant*, *p0-PUSCH-Alpha*, *P0-PUSCH-AlphaSet*, *SRI-PUSCH-PowerControl* as well as the SRI field in DCI format 0_0 and 0_1. The quantity P_O is provided as part of the power-control configuration and would typically depend on the target data rate but also on the noise and interference level experienced at the receiver.
- $N^{PUSCH}_{RB_{klm}}(i)$ denotes the bandwidth of PUSCH resource assignment expressed in number of resource blocks.
- $\alpha_{klm}(j)$ is a network-configurable parameter corresponding to fractional path loss compensation which is determined by RRC parameters *msg3-Alpha, ConfiguredGrantConfig, p0-PUSCH-Alpha, P0-PUSCH-AlphaSet, SRI-PUSCH-PowerControl* as well as the SRI field in DCI format 0_0 and 0_1. In the case of fractional path loss compensation ($\alpha < 1$), the path loss will not be fully compensated, and

the average received power will vary depending on the location of the device within the cell. In this case, the received power is lower for devices with higher path loss located at larger distances from the cell site. This must then be compensated by adjusting the uplink data rate. The advantage of fractional path loss compensation is reduced interference to neighboring cells, which is achieved at the expense of larger variation in the service quality, with reduced peak data-rate availability for devices closer to the cell edge.

- $PL_{klm}(q_d)$ is the downlink path loss estimate in dB calculated by the UE using reference signal index q_d for the active downlink BWP.
- $\Delta_{TF_{klm}}(i) = 10\log_{10}\left[(2^{BPRE \times K_s} - 1)\beta_{offset}^{PUSCH}\right]$ for $K_s = 1.25$ (provided by *deltaMCS*); otherwise $\Delta_{TF_{klm}}(i) = 0$ is related to the modulation scheme and channel coding rate used for the PUSCH transmission. The parameter bits per resource element (BPRE) is given by $BPRE = \sum_{r=0}^{C-1} K_r / N_{RE}$ for PUSCH with UL-SCH data and $BPRE = Q_m R / \beta_{offset}^{PUSCH}$ for CSI transmission in PUSCH without UL-SCH data wherein Q_m is the modulation order and R is the target code rate. This term models how the required received power varies when the number of information BPRE changes due to different modulation schemes and channel-coding rates.
- $f_{klm}(i,u)$ denotes PUSCH power control adjustment state which is given by $f_{klm}\left(i,u\right) = f_{klm}\left(i - i_0, u\right) + \sum_{v=0}^{C(D_i)-1} \delta_{PUSCH_{klm}}(v,u)$ wherein $\delta_{PUSCH}(.)$ is the power adjustment due to closed-loop power control. The power control commands can be sent to multiple devices by means of DCI format 2_2. Each power control command consists of 2 bits corresponding to four different update steps (-1, 0, $+1$, $+3$ dB). The reason for including 0 dB as an update step is that a power-control command is included in every scheduling grant, and it is desirable not to have to adjust the PUSCH transmit power for each grant.

  The PUCCH power control follows the same principles as PUSCH power control with some minor differences. For PUCCH power control, there is no fractional path loss compensation ($\alpha = 1$). Furthermore, for PUCCH power control, the closed-loop power control commands are carried within DCI formats 1_0 and 1_1, which are used for downlink scheduling assignments rather than within uplink scheduling grants. This is partly due to the fact that PUCCH transmission is used to carry HARQ-ACKs in response to a downlink transmission and such downlink transmissions are typically associated with downlink scheduling assignments on PDCCH and the corresponding power control commands could be used to adjust the PUCCH transmit power prior to the transmission of HARQ-ACKs. Similar to PUSCH, power control commands can also be carried jointly to multiple devices by means of DCI format 2_2.

  When the UE transmits PUCCH on active uplink BWP k of carrier l of serving cell m and PUCCH power control adjustment state with index u, it determines the PUCCH transmission power $P_{PUCCH_{klm}}(i, q_u, q_d, u)$ in the ith PUCCH transmission occasion as follows [8]:

$$P_{PUCCH_{klm}}(i, q_u, q_d, u) = \min\left\{P_{CMAX_{lm}}(i), P_{O\_PUCCH_{klm}}(q_u) + 10\log_{10}[2^{\mu}\mathrm{N}_{RB_{klm}}^{PUCCH}(i)] + PL_{klm}(q_d) + \Delta_{F\_PUCCH}(F) + \Delta_{TF_{klm}}(i) + g_{klm}(i,l)\right\}$$

where

- $P_{CMAX_{lm}}(i)$ is the configured UE transmit power.
- $P_{0\_PUCCH_{klm}}(q_u) = P_{0\_NOMINAL\_PUCCH} + P_{0\_UE\_PUCCH}(q_u)$ whose parameters are provided by *p0-nominal* and *p0-PUCCH-Value*.
- $N_{RB_{klm}}^{PUCCH}(i)$ is the bandwidth of PUCCH resource assignment expressed in number of resource blocks.
- $PL_{klm}(q_d)$ is the downlink path loss estimate in dB calculated by the UE using reference signal resource index q_d.
- $\Delta_{F\_PUCCH}(F)$ is PUCCH format-dependent power adjustment parameter provided by *deltaF-PUCCH-f0* for PUCCH format 0, *deltaF-PUCCH-f1* for PUCCH format 1, *deltaF-PUCCH-f2* for PUCCH format 2, *deltaF-PUCCH-f3* for PUCCH format 3, and *deltaF-PUCCH-f4* for PUCCH format 4.
- $\Delta TF_{klm}(i)$ is the PUCCH transmission power adjustment component which is dependent on the PUCCH format and the corresponding transport parameters.
- $g_{klm}(i, u)$ denotes the PUCCH power control adjustment state.

The UE calculates the transmission power for PRACH as $P_{PRACH_{klm}}(i) = \min(P_{CMAX_{lm}}(i), P_{PRACH\text{-}Target_{lm}} + PL_{klm})$ where $P_{CMAX_{lm}}(i)$ is the configured maximum UE transmission power, $P_{PRACH\text{-}Target_{lm}}$ is the target PRACH reception power corresponding to *PREAMBLE_RECEIVED_TARGET_POWER* parameter provided via RRC signaling, and PL_{klm} is path loss estimated for the active uplink BWP k of carrier l of serving cell m. The path loss is estimated based on the downlink reference signal associated with the PRACH transmission on the active downlink BWP of the mth serving cell and it is calculated by the UE as [*reference signal power*] − [*higher layer filtered RSRP*] in (dBm). If the active downlink BWP is the initial downlink BWP and for the SS/PBCH block and CORESET multiplexing pattern 2 or 3, the UE determines PL_{klm} based on the SS/PBCH block associated with the PRACH transmission [8].

The UE can set its configured maximum output power $P_{CMAX_{lm}}$ for the lth carrier of the mth serving cell in each slot. The configured maximum output power $P_{CMAX_{lm}}$ is set within $P_{CMAX-Low_{lm}} \le P_{CMAX_{lm}} \le P_{CMAX-High_{lm}}$ where $P_{CMAX-Low_{lm}} = \min\left[P_{EMAX_m}\Delta T_{C_m}, (P_{PowerClass} - \Delta P_{PowerClass}) - \max\left(MPR_m + A\text{-}MPR_m + \Delta T_{IB_m} + \Delta T_{C_m} + \Delta T_{RX_{SRS}}, P\text{-}MPR_m\right)\right]$ and $P_{CMAX-High_{lm}} = \min\left(P_{EMAX_m}, P_{PowerClass} - \Delta P_{PowerClass}\right)$. In the latter equation, P_{EMAX_m} is the value given by information element *P-Max* for the mth serving cell, and $P_{PowerClass}$ is the maximum UE power without taking into account the tolerance specified by 3GPP specifications [1,2].

The PRACH preamble transmission involves some uncertainty about the required transmit power. As a result, the PRACH preamble transmission includes a power-ramping

mechanism where the preamble may be retransmitted with a transmit power that is increased in steps during each transmission attempt. The device selects the initial PRACH preamble transmit power based on estimates of the downlink path loss in combination with a target received preamble power configured by the network. The path loss should be estimated based on the received power of the SS/PBCH block that the device has acquired and has determined the RACH resources for preamble transmission. If no random-access response is received within a predetermined window, the device can assume that the preamble was not correctly received by the network. In that case, the device repeats the preamble transmission with an increased transmit power. This power ramping continues until a random-access response has been received or until a configurable maximum number of retransmissions are attempted. Under such condition the random-access transmission has failed [14].

When uplink beamforming is used, the uplink path loss estimate $PL_{klm}(q_d)$ is used to determine the transmit power. The latter path loss estimate includes the effect of beamforming gain of the uplink beam pair to be used for PUSCH transmission. Assuming beam correspondence, this can be achieved by estimating the path loss based on measurements on a downlink reference signal transmitted over the corresponding downlink beam pair. As the uplink beam used for the transmission pair may change between PUSCH transmissions, the device may have to perform multiple path loss estimates corresponding to different candidate beam pairs. During PUSCH transmission over a specific beam pair, the path loss estimate corresponding to that beam pair is used to determine PUSCH transmit power. This is enabled by the parameter q in the path loss estimate $PL_{klm}(q_d)$[14].

The gNB configures the UE with a set of downlink reference signals based on which the path loss is estimated. Each reference signal is associated with a specific value of q. To limit the number of path loss estimations, the UE is not required to perform more than four parallel path loss estimations corresponding to different directions. The network configures a mapping between possible SRI values, provided in the scheduling grant, and different values of q. When a PUSCH transmission is scheduled by a scheduling grant including SRI, the path loss estimate associated with that SRI is used for determining the transmit power for the scheduled PUSCH transmission [14].

In the preceding equation for PUSCH power control, the open-loop parameters $P_{0\_PUSCH_{klm}}(j)$ and $\alpha_{klm}(j)$ are associated with parameter j, suggesting that there are multiple open-loop parameter pairs that can be used for different types of PUSCH transmissions, for example, Msg3 on PUSCH, grant-free PUSCH transmission, and scheduled PUSCH transmission. However, there is also a possibility to have multiple pairs of open-loop parameter for scheduled PUSCH transmission, where the pair to use for a certain PUSCH transmission can be selected based on the SRI similar to the selection of path loss estimates. In practice, this implies that the open-loop parameters $P_{0\_PUSCH_{klm}}(j)$ and $\alpha_{klm}(j)$ will depend on the uplink beam [14].

For PUSCH transmissions, the device can be configured with different open-loop parameter pairs $P_{0\_PUSCH_{klm}}(j)$ and $\alpha_{klm}(j)$ corresponding to different values of parameter j, where $j = 0$ is associated with Msg3 transmission and $j = 1$ is used in the case of grant-free PUSCH transmission. Each possible value of the SRI that can be provided as part of the uplink scheduling grant is associated with one of the configured open-loop parameter pairs. When a PUSCH transmission is scheduled with a certain SRI included in the scheduling grant, the open-loop parameters associated with that SRI are used when determining the transmit power for the scheduled PUSCH transmission [8].

The other parameter in PUSCH power control equation is the power control adjustment state with index u which is related to the closed-loop mechanism. PUSCH power control allows two independent closed-loop processes associated with $u = 0$ and $u = 1$. The u value(s) are provided by *sri-PUSCH-ClosedLoopIndex* RRC parameter. If PUSCH transmission is scheduled by a DCI format 0_1 and if DCI format 0_1 includes an SRI field, the UE determines the u value that is mapped to the SRI field. This means that similar to the case for multiple path loss estimates and multiple open-loop parameter pairs, the selection of u indicates the selection of the closed-loop process which is associated to the SRI included in the scheduling grant [8].

As we mentioned earlier, the UE relies on downlink measurements to calculate the path loss and to determine the power control parameters. The gNB determines the downlink transmit EPRE. For SS-RSRP, SS-RSRQ, and SS-SINR measurements, the UE may assume that downlink EPRE is constant across the bandwidth. The UE may further assume that downlink EPRE is constant over SSS carried in different SS/PBCH blocks, and the ratio of SSS EPRE to PBCH DM-RS EPRE is 0 dB [8]. For CSI-RSRP, CSI-RSRQ, and CSI-SINR measurements, the UE may assume downlink EPRE of a port of CSI-RS resource configuration is constant across the configured downlink bandwidth and constant across all configured OFDM symbols. The downlink SS/PBCH SSS EPRE can be derived from the SS/PBCH downlink transmit power given by the parameter *SS-PBCH-BlockPower* provided by RRC signaling. The downlink SSS transmit power is defined as the linear average over the power contributions (in Watts) of all REs that carry the SSS within the operating system bandwidth. The downlink CSI-RS EPRE can be derived from the SS/PBCH block downlink transmit power given by the parameter *SS-PBCH-BlockPower* and CSI-RS power offset given by the parameter *powerControlOffsetSS* provided through RRC signaling. The downlink reference signal transmit power is defined as the linear average over the power contributions [in (W)] of the resource elements that carry the configured CSI-RS within the operating system bandwidth.

For the purpose of downlink power allocation, the ratio of PDSCH EPRE to DM-RS EPRE ($\rho_{DM\text{-}RS}$ dB]) is given in Table 4.33 for downlink DM-RS associated with PDSCH, which depends on the number of DM-RS CDM groups without data [9]. It can be shown that the

Table 4.33: Ratio of PDSCH EPRE to DM-RS EPRE [9].

Number of DM-RS CDM Groups without Data	DM-RS Configuration Type 1 (dB)	DM-RS Configuration Type 2 (dB)
1	0	0
2	−3	−3
3	–	−4.77

Table 4.34: PT-RS EPRE to PDSCH EPRE per layer per resource element [9].

EPRE-Ratio $\rho_{PT\text{-}RS}$	Number of PDSCH Layers					
	1	2	3	4	5	6
0	0	3	4.77	6	7	7.78
1	0	0	0	0	0	0
2	Reserved					
3	Reserved					

DM-RS power scaling factor $\beta_{PDSCH}^{DM\text{-}RS}$ applied prior to DM-RS resource mapping is given by $\beta_{PDSCH}^{DM\text{-}RS} = 10^{(-\rho_{DM\text{-}RS}/20)}$.

When the UE is scheduled with a PT-RS port associated with the PDSCH, the ratio of PT-RS EPRE to PDSCH EPRE per layer per resource element for PT-RS port $\rho_{PT\text{-}RS}$ is given by Table 4.34. In that case, the PT-RS power scaling factor $\beta_{PT\text{-}RS}$ is given by $\beta_{PT\text{-}RS} = 10^{(\rho_{PT\text{-}RS}/20)}$; otherwise, it can be assumed that *epre-Ratio* is set to state "0" in Table 4.34.

References

3GPP *Specifications*[21]

[1] 3GPP TS 38.101-1, NR, User Equipment (UE) Radio Transmission and Reception; Part 1: Range 1 Standalone (Release 15), December 2018.

[2] 3GPP TS 38.101-2: NR, User Equipment (UE) Radio Transmission and Reception; Part 2: Range 2 Standalone (Release 15), December 2018.

[3] 3GPP TS 38.104, NR, Base Station (BS) Radio Transmission and Reception (Release 15), December 2018.

[4] 3GPP TS 38.133, NR, Requirements for Support of Radio Resource Management (Release 15), December 2018.

[5] 3GPP TS 38.202, NR, Services Provided by the Physical Layer (Release 15), December 2018.

[6] 3GPP TS 38.211, NR, Physical Channels and Modulation (Release 15), December 2018.

[21] 3GPP specifications can be accessed at the following URL: http://www.3gpp.org/ftp/Specs/archive/.

[7] 3GPP TS 38.212, NR, Multiplexing and Channel Coding (Release 15), December 2018.
[8] 3GPP TS 38.213, NR, Physical Layer Procedures for Control (Release 15), December 2018.
[9] 3GPP TS 38.214, NR, Physical Layer Procedures for Data (Release 15), December 2018.
[10] 3GPP TS 38.215, NR, Physical Layer Measurements (Release 15), December 2018.
[11] 3GPP TS 38.300, NR, NR and NG-RAN Overall Description, Stage 2 (Release 15), December 2018.
[12] 3GPP TS 38.321, NR, Medium Access Control (MAC) Protocol Specification (Release 15), December 2018.
[13] 3GPP TS 38.331, NR, Radio Resource Control (RRC), Protocol Specification (Release 15), December 2018.

Articles, Books, White Papers, and Application Notes

[14] E. Dahlman, S. Parkvall, 5G NR: The Next Generation Wireless Access Technology, Academic Press, August 2018.
[15] S. Ahmadi, LTE-Advanced: A Practical Systems Approach to Understanding 3GPP LTE Releases 10 and 11 Radio Access Technologies, Academic Press, November 2013.
[16] T.L. Marzetta, et al., Fundamentals of Massive MIMO, Cambridge University Press, December 2016.
[17] T.S. Rappaport, R.W. Heath Jr, Millimeter Wave Wireless Communications, Prentice Hall, September, 2014.
[18] S. Lin, D.J. Costello, Error Control Coding, second ed., Prentice Hall, June 2004.
[19] C.E. Shannon, W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, September 1949.
[20] A. Papoulis, Probability, Random Variables and Stochastic Processes, fourth ed., McGraw Hill Higher Education, January 2002.
[21] T.M. Cover, J.A. Thomas, Elements of Information Theory, John Wiley & Sons, July 2006.
[22] J.G. Proakis, Digital Communications, fifth ed., McGraw-Hill, November 2007.
[23] F. Ademaj, et al., 3GPP 3D MIMO channel model: a holistic implementation guideline for open source simulation tools, EURASIP J. Wirel. Commun. Netw. 2016 (2016) 55.
[24] R.G. Gallager, Principles of Digital Communication, Cambridge University Press, March 2008.
[25] B. Mondal, et al., 3D channel model in 3GPP, IEEE Commun. Mag. 53 (3) (2015).
[26] F. Vook, 3GPP new radio at sub-6GHz: features and performance, in: IWPC Workshop on 5G NR Mobile Networks and User Equipment, January 2018.
[27] T.S. Rappaport, et al., Overview of millimeter wave communications for fifth-generation (5G) wireless networks—with a focus on propagation models, in: IEEE Transactions on Antennas and Propagation, Special Issue on 5G, November 2017.
[28] S. Cammerer, Spatially Coupled LDPC Codes, Institut für Nachrichtenübertragung. WebDemos. Available from: <http://www.inue.uni-stuttgart.de/lehre/demo.html>.
[29] A. Elkelesh, M. Ebada, Polar Codes, Institut für Nachrichtenübertragung. WebDemos. Available from: <http://www.inue.uni-stuttgart.de/lehre/demo.html>.
[30] 5G New Radio, ShareTechNote. Available from: <http://www.sharetechnote.com>.
[31] J. Campos, Understanding the 5G NR Physical Layer, Keysight Technologies, November 2017.
[32] MediaTek, 5G NR: A New Era for Enhanced Mobile Broadband, White paper, March 2018.
[33] E. Arikan, Channel polarization: a method for constructing capacity achieving codes for symmetric binary-input memoryless channels, IEEE Trans. Inf. Theory 55 (7) (2009) 3051−3073.
[34] E. Arikan, Systematic polar coding, IEEE Commun. Lett. 15 (8) (2011) 860−862.
[35] R.G. Maunder, The implementation challenges of polar codes, in: AccelerComm White Paper, February 2018.
[36] V. Bioglio, C. Condo, I. Land, Design of Polar Codes in 5G New Radio, Cornell University Library, 2019.
[37] F. Hamidi-Sepehr, et al., Analysis of 5G LDPC codes rate-matching design, in: IEEE Vehicular Technology Conference (VTC), June 2018.

[38] 3GPP TSG RAN WG1, R1-1608862, Polar Code Construction for NR, Huawei, HiSilicon, 2016.
[39] R.G. Gallager, MIT Press Classic Series Low-Density Parity-Check Codes, MIT Press, Cambridge, MA, 1963.
[40] F. Sabatier, Polar Coding Tutorial. Available from: <http://ipgdemos.epfl.ch/polarcodestutorial/index.html>.
[41] C. Berrou, A. Glavieux, P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: turbo-codes, in: Proceedings of IEEE International Communications Conference (ICC), May 1993.
[42] M. Giordani, et al., A tutorial on beam management for 3GPP NR at mmWave frequencies, IEEE Commun. Surv. Tutorials (2018).
[43] H. Shariatmadari, et al., Fifth-generation control channel design, achieving ultra-reliable low-latency communications, IEEE Veh. Technol. Mag. 13 (2) (2018) 84–93.
[44] 5G New Radio: Introduction to the Physical Layer, White Paper, National Instruments, 2018.
[45] S.M. Razavizadeh, et al., Three-dimensional beamforming: a new enabling technology for 5G wireless networks, IEEE Signal Process Mag. 31 (6) (2014) 94–101.
[46] H. Ji, et al., Overview of full-dimension MIMO in LTE-Advanced Pro, IEEE Commun. Mag. 55 (2) (2017) 176–184.
[47] Y.-H. Nam, et al., Full dimension MIMO for LTE-advanced and 5G, in: Information Theory and Applications Workshop (ITA), February 2015.
[48] A. Alkhateeb, G. Leus, R.W. Heath, Multi-layer precoding: a potential solution for full-dimensional massive MIMO systems, IEEE Trans. Wireless Commun. 16 (9) (2017) 5810–5824.
[49] J. Jeon, NR wide bandwidth operations, IEEE Commun. Mag. 56 (3) (2018) 42–46.
[50] D.H.N. Nguyen, T. Le-Ngoc, MMSE precoding for multiuser MISO downlink transmission with non-homogeneous user SNR conditions, EURASIP J. Adv. Signal Process. 2014 (2014) 85.
[51] R.-A. Pitaval, et al., Overcoming 5G PRACH capacity shortfall by combining Zadoff-Chu and M-sequences, IEEE Int. Conf. Commun. (ICC) (2018).
[52] T.L. Marzetta, Massive MIMO: an introduction, Bell Labs Tech. J. 20 (2015) 11–22.
[53] F. Hamidi-Sepehr, et al., 5G NR PDCCH: design and performance, in: 2018 IEEE 5G World Forum (5GWF), July 2018.
[54] Y. Qi, et al., On the Phase Tracking Reference Signal (PT-RS) Design for 5G New Radio (NR), Cornell University Library, 2018.
[55] Y. Blankenship, et al., Channel coding in 5G new radio: a tutorial overview and performance comparison with 4G LTE, IEEE Veh. Technol. Mag. 13 (4) (2018) 60–69.
[56] M. Reil, G. Lloyd, Millimeter-wave beamforming: antenna arrays & characterization, White Paper, Rohde & Schwarz, 2016.
[57] E. Onggosanusi, et al., Modular and high-resolution channel state information and beam management for 5G new radio, IEEE Commun. Mag. 56 (3) (2018) 48–55.
[58] A.L. Swindlehurst, et al., Millimeter-wave massive MIMO: the next wireless revolution? IEEE Commun. Mag. 52 (9) (2014) 56–62.
[59] Y. Huang, et al., Multi-panel MIMO in 5G, IEEE Commun. Mag. 56 (3) (2018) 56–61.
[60] T.L. Marzetta, Noncooperative cellular wireless with unlimited numbers of base station antennas, IEEE Trans. Wireless Commun. 9 (11) (2010) 3590–3600.
[61] F. Rusek, et al., Scaling up MIMO: opportunities and challenges with very large arrays, IEEE Signal Process Mag. 30 (1) (2013) 40–60.
[62] E.G. Larsson, et al., Massive MIMO for next generation wireless systems, IEEE Commun. Mag. 52 (2) (2014) 186–195.
[63] L. Lu, et al., An overview of massive MIMO: benefits and challenges, IEEE J. Sel. Top. Signal Process. 8 (5) (2014) 742–758.
[64] W. Roh, et al., Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results, IEEE Commun. Mag. 52 (2) (2014) 106–113.

[65] 3GPP TSG RAN WG1, R1-1800036, Summary of Remaining Issues on HARQ Management, Huawei, HiSilicon, 2018.
[66] 3GPP TSG RAN WG1, R1-1712546, Discussion on Codebook Subset Restriction for NR, Intel Corporation, 2017.
[67] 5G Toolbox, The MathWorks, Inc. Available from: <https://www.mathworks.com/help/5g/index.html>.
[68] H. Koorapaty, NR physical layer design: physical layer structure, numerology and frame structure, in: RWS-180007, Workshop on 3GPP Submission Towards IMT-2020, October 2018.
[69] Y. Kim, NR physical layer design: NR MIMO, in: RWS-180008, Workshop on 3GPP Submission Towards IMT-2020, October 2018.
[70] G. Noh, et al., DM-RS design and evaluation for 3GPP 5G new radio in a high speed train scenario, IEEE Global Commun. Conf, 2017.
[71] X. Lin, et al., 5G New Radio: Unveiling the Essentials of the Next Generation Wireless Access Technology, Cornell University Library, 2018.
[72] J. Liu, et al., Initial access, mobility, and user-centric multi-beam operation in 5G new radio, IEEE Commun. Mag. 56 (3) (2018) 35–41.
[73] Updated information on China's IMT-2020 Submission, IEEE ComSoc Technology Blog, October 2018 (https://techblog.comsoc.org/2018/10/19/updated-information-on-chinas-imt-2020-submission/).
[74] W.U. Yong, Self evaluation: enhanced mobile broadband (eMBB) evaluation results, in: RWS-180018, Workshop on 3GPP Submission Towards IMT-2020, October 2018.
[75] D.J. Love, R.W. Heath Jr., Grassmannian precoding for spatial multiplexing systems, in: Proceedings of the Allerton Conference on Communication Control and Computing, Monticello, 2003.
[76] N. Jindal, MIMO broadcast channels with finite-rate feedback, IEEE Trans. Inf. Theory 52 (11) (2006).