CHAPTER 8

Multiple-input multiple-output enhancement and beam management

Zhihua Shi, Wenhong Chen, Yingpei Huang, Jiejiao Tian, Yun Fang, Xin You and Li Guo

Starting from 4G LTE, multiple-input multiple-output (MIMO) transmission has been one of the key technologies that continuously improves the spectrum efficiency by exploiting the freedom in the spatial-domain. In NR, the data can be transmitted in high-frequency band (e.g., millimeter wave), which leads to new challenges in MIMO technology. On the one hand, the bandwidth in high-frequency band is larger and an antenna array can consist of a larger number of antenna elements due to the smaller antenna size in the high-frequency band (i.e., Massive MIMO). However, the path loss is greater, and the coverage distance is consequently smaller. On the other hand, with the increase of the number of antennas, the complexity of MIMO processing may increase exponentially, and it is harder to implement many MIMO transmission schemes (such as massive digital precoding and spatial-multiplexing) in this scenario. In order to address the abovementioned issues, the NR adopted analog beam-forming as a key feature to overcome the limitations on complexity and coverage. Through using narrower beam, larger beam-forming gain can be obtained. Moreover, the complexity of analog beam will be much lower than the traditional MIMO schemes based on fully digital processing.

In order to support new scenarios, and also to improve the performance in frequency bands below 6 GHz, the NR introduced several enhancements to the MIMO technologies: more refined and more flexible channel state information (CSI) feedback, beam management, beam failure recovery (BFR), enhanced reference signal (RS), and multiple transmission and reception point (multi-TRP) cooperative transmission. Among them, the RS design in NR basically reuses the design principles of LTE with some optimization, which will not be introduced in detail

5G NR and Enhancements
DOI: https://doi.org/10.1016/B978-0-323-91060-6.00008-8

here. This chapter will focus on the new features of NR that are not supported in LTE, including CSI feedback enhancement (e.g., R15 Type II codebook), beam management and BFR introduced in R15/R16, enhanced eType II codebook in R16, and multi-TRP transmission.

8.1 CSI feedback for NR MIMO enhancement

CSI feedback is an important part of NR MIMO enhancement. In order to support a large number of antenna ports, more flexible CSI reporting, and more accurate CSI, the NR introduced various enhancements for CSI feedback, such as configuration signaling, triggering scheme, measurement method, codebook design, UE capability reporting, etc. This section will introduce the enhancements and optimizations made in NR on the basis of the LTE CSI feedback mechanism, and the Type II codebook that is newly introduced in R15 will be the focus of the discussion.

8.1.1 CSI feedback enhancement in NR

8.1.1.1 Enhancement on CSI configuration signaling

The overall framework of NR CSI feedback is established on the basis of the CSI feedback mechanism of the LTE system. In the initial framework of CSI configuration [1] agreed on at the beginning of NR discussion, CSI-related resource allocation and reporting configuration were supported through the following four parameter sets:

- N CSI reporting configurations are used to configure the resources and modes of CSI reporting, similar to the CSI process of LTE.
- M channel measurement resource configurations are used to configure the RS for channel measurement.
- J interference measurement resource (IMR) configurations are used to configure RSs resources for interference measurement, similar to IMR in LTE.
- One CSI measurement configuration indicates the association between the above N CSI reporting configurations, M channel measurement resource configurations and J IMR configurations.

At the RAN1#AdHoc1701 meeting, the abovementioned channel measurement resource configuration and IMR configuration are combined into a CSI resource configuration [2]. It was also clarified that the CSI measurement configurations can contain L association indications, each of which is used to associate a CSI reporting configuration with a CSI resource configuration. However, RAN2 does not fully adopt the framework designed by RAN1

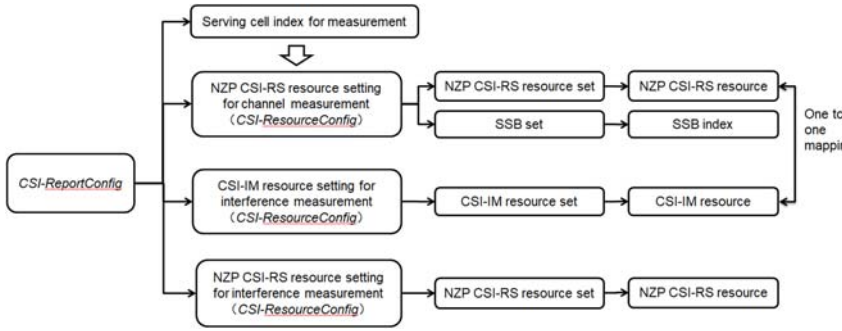


Figure 8.1 CSI resource configuration of NR.

for designing RRC signaling, and instead a few CSI resource configurations for CSI measurement are explicitly included in the CSI reporting configuration, thereby avoiding the additional signaling needed to indicate the relationship between CSI reporting configuration and CSI resource configuration. The UE can obtain all the configuration information needed for CSI measurement and reporting directly from a CSI reporting configuration. Each CSI resource configuration can contain multiple nonzero-power CSI RS (NZP CSI−RS) resource sets for channel measurement (where each set can contain several CSI−RS resources corresponding to the above RS configuration), and multiple CSI−RS resource sets for interference measurement (CSI−IM) (where each set can contain multiple resources for interference measurement, corresponding to the interference measurement configuration mentioned above). For aperiodic CSI reporting, a CSI resource configuration can further include multiple NZP CSI−RS resource sets for interference measurement (as shown in Fig. 8.1).

8.1.1.2 Enhancement on interference measurement

During the discussion of CSI feedback, there were three candidate schemes for interference measurement:

- Similar to LTE, IMR based on zero-power CSI−RS is used to measure the intercell interference.
- The interference measurement is based on the NZP CSI−RS resources [3]. In this scheme, the network adopts the prescheduling method to precode the CSI−RS port(s), so that the UE can measure the interference of multiplexed users based on the precoded CSI−RS port(s). Then the network side can use the measurement results for subsequent data scheduling.

- The interference measurement is based on the demodulation RS (DMRS) port(s) [4]. In this scheme, the UE estimates the possible interference of multiplexed users on the unused DMRS port(s).

Among them, the first scheme follows the LTE method and was agreed on at the early stage of NR standardization. The latter two schemes have similar effects, and they support the same interference type. The advantage of the scheme using NZP CSI−RS is that the interference of each multiplexing port in subsequent scheduling can be estimated accurately, while its disadvantage is that it requires additional resources needed for interference measurement and increases the measurement complexity. The DMRS-based scheme does not require additional resources, and can guarantee high-measurement accuracy in case of continuous PDSCH scheduling. However, this method is only able to measure the channel in the current scheduled bandwidth. In practical systems, it is difficult for the physical resources, the paired users, and the precoding matrix of the two continuous scheduling to remain consistent. After discussions in several meetings, the interference measurement scheme based on NZP CSI−RS was finally agreed on at the RAN1#89 meeting [5], and DMRS-based interference measurement was excluded. In the subsequent standardization process, interference measurement based on NZP CSI−RS resources was further limited to be only used for aperiodic CSI reporting (Fig. 8.2).

8.1.1.3 Enhancement on reporting periodicity

Both periodic and aperiodic CSI reporting are supported in the LTE system. In addition to these two methods, NR supports semipersistent CSI

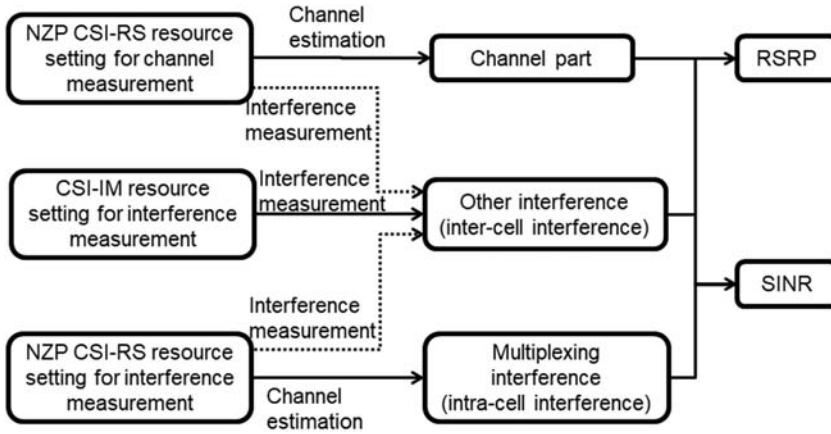


Figure 8.2 Channel and interference measurement mechanism in NR.

reporting, and also further enhances the mechanism of aperiodic CSI reporting. Specifically, NR supports the following three CSI reporting schemes:

- The periodic CSI reporting follows the LTE method, and RRC signaling is used to indicate the configuration of CSI reporting.
- Semipersistent CSI reporting can be implemented via two different ways: the method of PUSCH reporting based on DCI scheduling and the method of PUCCH reporting based on MAC CE activation. The necessity and benefits of the former method was extensively discussed in 3GPP. Finally, both methods were adopted in NR as a compromise. The specific choice depends on the UE capability and network side configuration.
- For aperiodic CSI reporting, two different triggering schemes were discussed: the method of RRC + DCI and the method of RRC + MAC + DCI. The latter allows higher configuration flexibility, but also introduces additional signaling design. Since both schemes are supported many companies, a compromise solution was agreed on in the last two RAN1 meetings of R15: when the number of aperiodic CSI trigger states configured by RRC exceeds the number that can be indicated by CSI trigger field in DCI, MAC CE is used to activate some of the states to be triggered through DCI. In this way, whether to use MAC CE for activation completely depends on the configuration on the network side. In addition, if the measurement of aperiodic CSI is based on aperiodic measurement resources (such as aperiodic CSI−RS), the corresponding trigger signaling also indicates the transmission of the aperiodic RS used for measurement and the reporting of aperiodic CSI (as shown in Fig. 8.3).

On the basis of the PUSCH-based aperiodic CSI reporting in NR that is similar to that in LTE, in the RAN1#AdHoc1709 meeting, 3GPP agreed to support DCI-triggered PUCCH-based aperiodic CSI reporting [6]. Some companies thought the PUCCH-based method required high implementation complexity and great standardization effort, and thus it is difficult to complete it within R15 time window. In the RAN1#90bis meeting, three candidate trigger methods were discussed for PUCCH-based aperiodic CSI reporting [7]: triggered by DCI scheduling PDSCH, triggered by DCI scheduling PUSCH (RRC configures PUCCH or PUSCH), and triggered by DCI scheduling PUSCH (DCI configures PUCCH or PUSCH), and conclusion was reached. During the email discussion after the meeting, there was still great divergence among

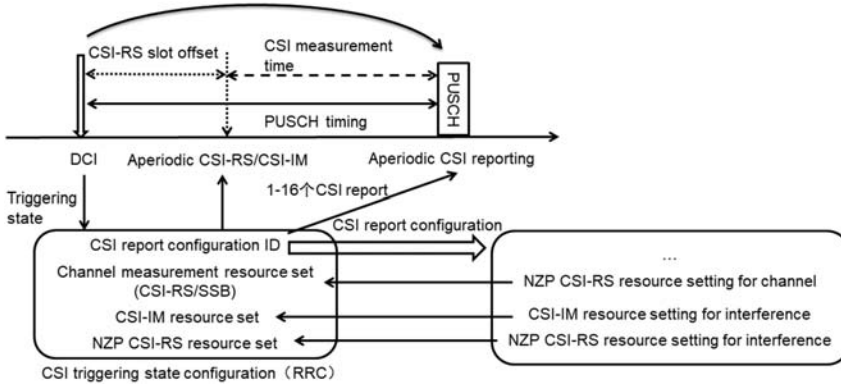


Figure 8.3 Aperiodic CSI−RS triggering and aperiodic CSI reporting mechanism in NR.

companies on which DCI is used to trigger the PUCCH-based CSI reporting. In the end, the solution of PUCCH-based aperiodic reporting was not supported in NR standardization.

8.1.1.4 Enhancement on content of CSI reporting

In addition to the CSI reporting method, the NR has also enhanced the content of CSI reporting so that the CSI feedback in the NR can support more functions. The CSI report in the LTE system can only support CSI for PDSCH scheduling, such as CSI−RS resource indicator (CRI), rank indicator (RI), precoding matrix indicator (PMI), and channel quality indicator (CQI). The NR introduces beam management-related content in CSI, such as CRI/SSB index and corresponding RS reception power (RSRP). The network side can determine the downlink beam according to the feedback from the UE. Furthermore, a CSI reporting configuration in NR can also correspond to a CSI report with empty content (i.e., the UE does not need to perform actual CSI reporting). Such a CSI reporting configuration can be used to trigger aperiodic tracking RS (TRS) used for time-frequency synchronization or aperiodic CSI−RS used in downlink beam management. The UE does not need to perform reporting based on these RSs (Table 8.1).

8.1.1.5 Enhancement of reciprocity-based CSI

In the LTE system, the UE can report the corresponding CQI based on nonprecoded CSI−RS ports and a predefined transmission scheme. Other CSI information is derived by the network side based on channel reciprocity. In order to better support downlink transmission based on channel

Table 8.1 CSI reporting contents and corresponding application scenarios in NR.

Report content	Objective	Applicability	Notes
"cri−rI−PMI−CQI"	CSI reporting based on PMI	Type I/II codebook	It can support CSI−RS-based beam selection and CSI reporting, similar to LTE class A/B CSI reporting.
"cri−rI−LI−PMI−CQI"	CSI reporting based on PMI	Type I/II codebook	The strongest transport layer indicator LI is added to determine the transport layer associated with DL PTRS.
"cri−rI−i1"	Partial CSI report	Type I codebook	Only wideband channel information is reported, which can be used together with other reported content.
"cri−rI−i1-CQI"	CSI report for semiopen-loop transmission	Type I codebook	Only W_1 is reported and W_2 is not reported. The UE randomly selects one W_2 corresponding to W_1 on each PRG to calculate CQI.
"cri−rI−CQI"	No-PMI-based CSI reporting	indicated CSI−RS port	RI and CQI are estimated based on CSI−RS ports indicated for each rank from network devices.
"cri−rSRP"	Tx beam management	RSRP measurement	RSRP measurement and reporting based on CSI−RS resources (with different beams).
"ssb-Index−rSRP"	Tx beam management	RSRP measurement	RSRP measurement and reporting based on SSB (with different beams).
"none"	TRS or downlink receive beam management	CSI reporting based on PUSCH	It cannot be configured for CSI reporting on PUCCH.

reciprocity, at the RAN1#88bis meeting, a non-PMI-based feedback solution was agreed on for the case that the network can obtain complete downlink channel information from uplink measurement. In this solution, the CSI contains RI and CQI, where CQI is calculated based on a codebook, and precoding information is derived based on channel reciprocity and the reported RI. Regarding the codebook assumed for CQI feedback, three candidate schemes were agreed on at the RAN1#90 meeting [8]: using port selection codebook, using columns of an identity matrix as the codebook, or using the existing codebook. In the next RAN1 meeting, the third scheme was first excluded, and the first two schemes further evolved into two port selection methods:

- Method 1: The CQI is calculated based on the port selection codebook, and each column is used to select a CSI−RS port corresponding to a layer.
- Method 2: Port index indicated by the network is used to select part of CSI−RS ports from the CSI−RS ports in a CSI−RS resource for RI/CQI calculation. The network indicates the CSI−RS ports used for measurement for each rank.

At the RAN1#90bis meeting, the latter was agreed upon. The network first determines the candidate precoding vectors based on the channel reciprocity, and CSI−RS ports are precoded based on the determined precoding vectors. Then the network informs the UE with the CSI−RS ports for measurement corresponding to different ranks in advance. The UE reports RI and CQI based on measuring the indicated CSI−RS ports, where the UE performs CQI calculation based on the CSI−RS ports corresponding to the reported rank. The CSI feedback mechanism for the case of partial channel reciprocity (i.e., the UE has fewer transmit antennas than reception antennas, and the network can only obtain part of the downlink channel information) was also discussed during the standardization process. Companies performed a lot of evaluation based on the candidate feedback mechanisms for that, but no scheme was finally agreed on due to the large divergence among companies.

8.1.2 R15 Type I codebook

R15 supports two types of codebooks: Type I codebook and Type II codebook. R16 further supports the enhanced eType II codebook. The design of Type I codebook basically follows the codebook design of LTE, and it supports normal spatial-resolution and CSI accuracy. Type II and

eType II codebooks adopt the design idea of eigenvector quantization and eigenvector linear weighting to support higher spatial-resolution and CSI accuracy. This section will focus on the design of the Type I codebook, while the design of the Type II and eType II codebooks will be discussed in the following chapters.

At the initial stage of the NR discussion, it was agreed that the NR Type I codebook could reuse the two-stage codebook design in the LTE system (i.e., $\boldsymbol{W} = \boldsymbol{W}_1\boldsymbol{W}_2$,) where $\boldsymbol{W}_1$ is used to report a beam (or beam group), and $\boldsymbol{W}_2$ is used to report at least one of the following information: the beam selected from the beam group, the weighting coefficient between the beams, and the phase between polarization directions. Two issues need to be solved for designs based on such a scheme: whether the beam is selected only based on $\boldsymbol{W}_1$ (number of beams in a beam set $L = 1$) or based on both $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ as LTE ($L = 4$) and whether a single beam or a linear weighting of multiple beams is applied to a layer. It should be noted that the beams used in codebook design are generally digital beams that are mapped to CSI−RS antenna ports (i.e., the beams formed by digital precoding). They are different from the analog beams discussed in other chapters.

At the RAN1#89 meeting, more than 30 companies jointly presented a package proposal for NR codebook design [9], which provided most of the designs for Type I and Type II codebooks, including the solutions for the above two issues. Although some details of the design were still controversial and not acceptable to some companies, the joint proposal was finally approved due to majority supporting. According to that proposal, with the exception of the 2-port codebook that reuses the design of LTE, all other codewords in Type I codebook can be expressed as: $\boldsymbol{W} = \boldsymbol{W}_1\boldsymbol{W}_2$, where $\boldsymbol{W}_1 = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{bmatrix}$ and $\boldsymbol{B} = [\boldsymbol{b}_0, \boldsymbol{b}_1 \ldots, \boldsymbol{b}_{L-1}]$ corresponds to L oversampled DFT beams (they can be horizontal and vertical two-dimensional beams). Although adopting similar expressions, the NR introduces many enhancements in the Type I codebook compared to LTE:

- When rank = 1 or 2, the network can configure either $L = 1$ or $L = 4$. If $L = 1$, $\boldsymbol{W}_2$ is used to only reports the phase between two polarizations. If $L = 4$, $\boldsymbol{W}_2$ is used to select a beam from the beam group corresponding to $\boldsymbol{W}_1$ and also report the interpolarization phase. The definition of a beam group in NR reuses one of the three definitions of beam groups in LTE, as shown in Fig. 8.4.

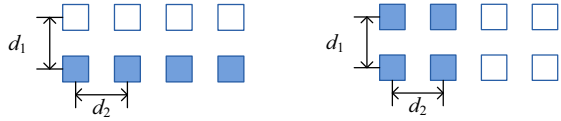


Figure 8.4 Pattern of beam group supported by $L = 4$ (*left*: horizontal port, *right*: 2D port).

- When rank = 3 or 4, only $L = 1$ is supported. The LTE codebook design is reused when the number of ports is less than 16, while the phase between orthogonal beams is additionally reported for 16 and 32 ports.
- When rank is greater than 4, only $L = 1$ is supported and fixed orthogonal beams are used.
- Based on the single-panel codebook, the multipanel codebook is designed with an additional reporting interpanel phase (which can be wideband or subband). A single-panel codebook is adopted for each panel.
- The codebook subset restriction (CSR) is introduced for the Type I codebook. The CSR can be performed for each DFT beam and each rank, and the PMI corresponding to the restricted beam is not reported by the UE.

Based on the single-panel codebook, CSI reporting for semiopen-loop transmission is also supported in NR, where the UE only reports CRI/RI and beam information ($\boldsymbol{W}_1$), while $\boldsymbol{W}_2$ is not reported. The UE assumes that network adopts the beam(s) corresponding to $\boldsymbol{W}_1$ for downlink transmission. There are many candidate schemes for the precoding assumption that is used by the UE to calculate CQI, which are diversity transmission (SFBC, similar to CQI assumption used in LTE transmission mode 7) and codeword polling and random selection of codewords. At the RAN1AdHoc1709 meeting, it was agreed that the UE assumes the codeword used in each precoding resource group (PRG) is randomly selected from the multiple $\boldsymbol{W}_2$ corresponding to the reported $\boldsymbol{W}_1$. The codewords for random selection can be further indicated via the CSR configured by the network.

8.1.3 R15 Type II codebook

The previous section introduced the Type I codebook design in NR. It can meet the requirement of low-precision CSI feedback with lower feedback overhead for SU−MIMO. However, in some application

scenarios such as MU−MIMO, CSI feedback with higher channel precision is desired. For this reason, RAN1# 86bis agreed that the Type II codebook can support high-precision CSI in NR [10].

The main differences between Type I and Type II codebooks are:

- The Type I codebook is mainly used for SU−MIMO and can support high-rank transmission. The Type II codebook is mainly for MU−MIMO and rank is generally low. Specifically, only rank = 1/2 is supported in the Type II codebook for low overhead.
- The Type I codebook reports only one beam per precoder, while the Type II codebook supports reporting combinations of multiple beams.
- The power of the Type I codebook ports is constant, while power on each port of the Type II codebook varies due to superposition.
- The layers in the Type I codebook are orthogonal to each layer, while the Type II codebook does not have such a constraint.
- Only phase information is reported on the subband in the Type I codebook, while the amplitude coefficient of each subband can be reported in the Type II codebook through the mode of wideband + subband.
- The feedback payload in the Type I codebook is low, only dozens of bits, which can be obtained by exhaustive searching. However, the feedback of the Type II codebook requires hundreds of bits (∼500 bits) and the weighting coefficients are computed by minimizing the mean square error and then quantization.
- The feedback payload of the Type I codebook is fixed, while the feedback payload of the Type II codebook varies with the channel state.

8.1.3.1 Structure of the Type II codebook

At the RAN1#AdHoc1701 meeting, it was agreed to support high-precision CSI reporting through a linear combination of multiple beams [11]. The spatial-domain (SD) channel information is projected onto a set of bases, and then the UE reports dominated components and the weighting coefficients. There exist three different candidate schemes: channel correlation matrix feedback, precoding matrix feedback, and hybrid feedback. The method of precoding matrix feedback extends the design of LTE. The UE recommends reporting precoder and RI/CQI information and determines the transmission rate associated with the precoder. In the method of correlation matrix feedback, the UE reports the long-term and wideband channel covariance matrix. The method of hybrid feedback is

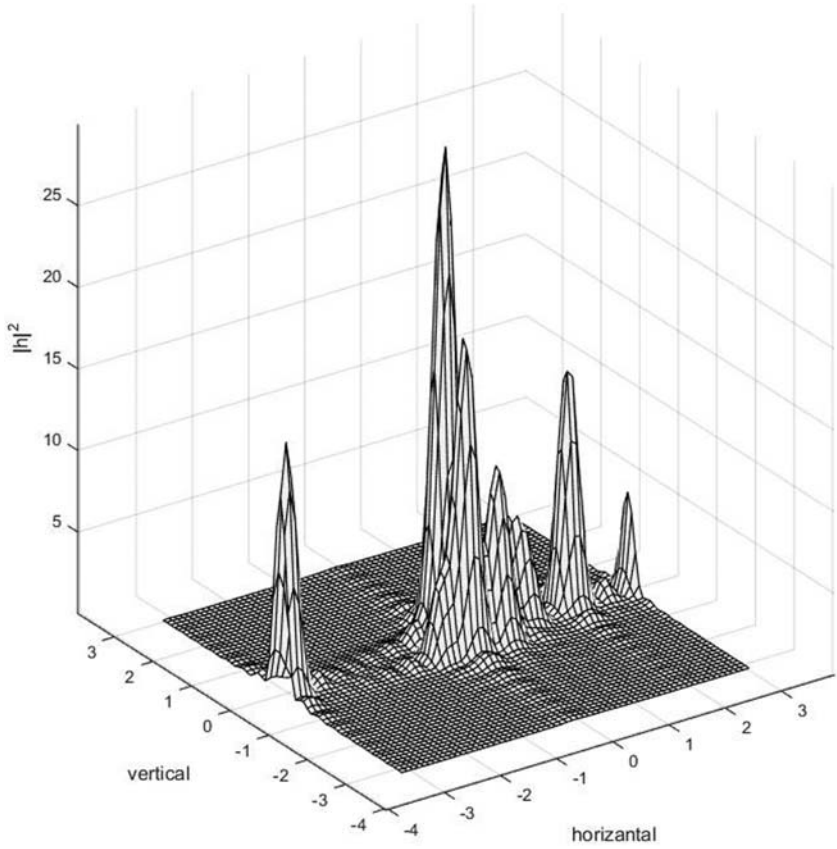


Figure 8.5 Sparsity of channel in angular domain.

similar to Class B in LTE, and the CSI−RS port with beam-forming is selected for feedback.

The basic scheme of linear combining [12−15] is shown in Fig. 8.5. The principle is to transform the spatial-channel into the angle domain, and the combining coefficient represents the amplitude and phase per component. On the one hand, the radio channel itself is sparse in space, which means only few directions have energy. When spatial-sampling rate increases with more antenna, only a few coefficients have nonzero energy. The UE only needs to feed back a few nonzero components, which greatly reduces the reporting overhead. On the other hand, each precoder (e.g., eigenvectors) can also be expressed as a linear combination of channel spatial-vectors (Fig. 8.6).

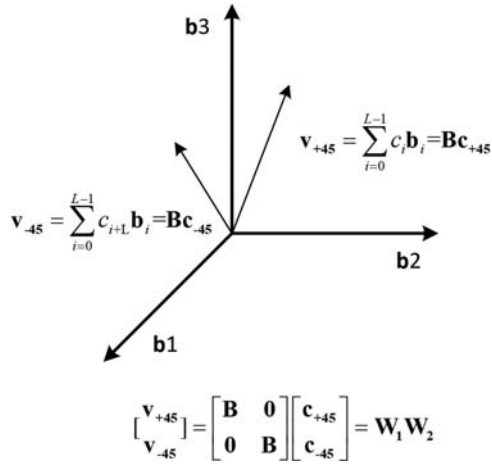


Figure 8.6 Schematic diagram of linear combination.

Regarding the codebook structure, most companies agreed to adopt the dual codebook structure for the Type II codebook [12]:

$$\boldsymbol{W} = \boldsymbol{W}_1 \boldsymbol{W}_2 \tag{8.1}$$

At the RAN1#AdHoc1701 meeting, the candidate schemes for the Type II codebook were agreed on. Similar as Type I codebook, two models are discussed. The first model is:

$$\boldsymbol{W}_1 = \begin{bmatrix} \boldsymbol{B}_1 & \boldsymbol{B}_2 \\ \boldsymbol{B}_1 & -\boldsymbol{B}_2 \end{bmatrix} \tag{8.2}$$

And another one is (Fig. 8.6):

$$\boldsymbol{W}_1 = \begin{bmatrix} \boldsymbol{B} & 0 \\ 0 & \boldsymbol{B} \end{bmatrix} \tag{8.3}$$

The above two alternatives are actually equivalent. There exist two options for the selection of vector $\boldsymbol{B}$. The first one ensures $\boldsymbol{B}$ to be a set of orthogonal vectors, while the second option allows $\boldsymbol{B}$ to be nonorthogonal. Finally, at the RAN1#89 meeting [16], most companies agree to use the same beam for both polarizations, thus $\boldsymbol{W}_1$ has a block diagonal structure.

In order to improve the feedback accuracy, NR adopts four times as much oversampled 2D−DFT vectors (Fig. 8.7) to quantize the beam in both horizontal and vertical directions. Similar to the codebook

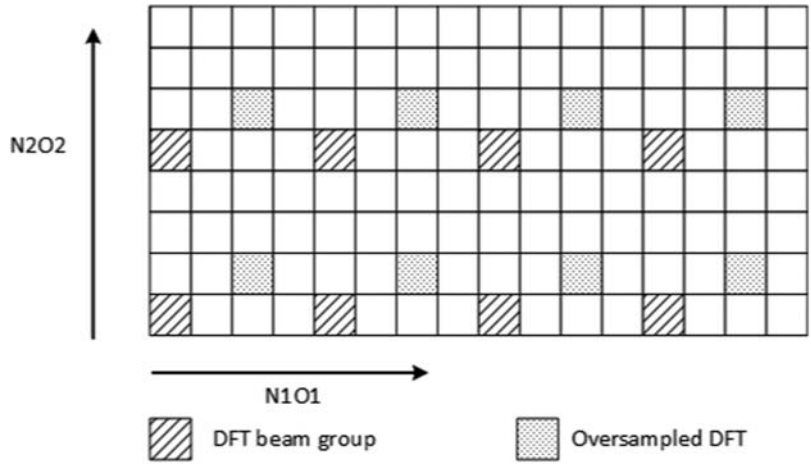


Figure 8.7 Diagram of beam selection.

Table 8.2 Codebook configuration of Type II codebook.

CSI−RS port	(N_1, N_2)	(O_1, O_2)
4	(2,1)	(4,-)
8	(2,2)	(4,4)
	(4,1)	(4,-)
12	(3,2)	(4,4)
	(6,1)	(4,-)
16	(4,2)	(4,4)
	(8,1)	(4,-)
24	(6,2), (4,3)	(4,4)
	(12,1)	(4,-)
32	(8,2), (4,4)	(4,4)
	(16,1)	(4,-)

introduced in LTE R14, the orthogonality of L selected beams is considered, where $L \leq N_1 N_2$. The number of CSI−RS ports supported by the Type II is shown in Table 8.2.

In the above expression, $\boldsymbol{W}_2$ is the combined coefficients corresponding to L beams on the subband, including the amplitude and phase. The coefficients per-layer and per-polarization direction are selected independently. The codebook can be expressed as:

$$\tilde{\boldsymbol{w}}_{r,l} = \sum_{i=0}^{L-1} \boldsymbol{b}_{k_1^{(i)} k_2^{(i)}} \times p_{r,l,i}^{(WB)} \times p_{r,l,i}^{(SB)} \times c_{r,l,i} \tag{8.4}$$

where $\boldsymbol{b}_{k_1^{(i)}k_2^{(i)}}$ is an oversampled DFT vector, which is determined by the $\boldsymbol{W}_1$ vector; $p_{r,l,i}^{(WB)}$ is the wideband amplitude of the ith beam in the polarization r and layer l; $p_{r,l,i}^{(SB)}$ is the subband amplitude of the ith beam in the polarization r and layer l; $c_{r,l,i}$ is the phase of the ith beam in the polarization r and layer l.

8.1.3.2 Quantization

During the 3GPP discussion, some companies believed that subband reporting can bring better trade-off between overhead and performance. Wideband reporting has low overhead while its performance is slightly worse than that of the subband reporting. In subband reporting, subband amplitude is reported in differential manner [17]. The feedback payload increases but the system performance is also improved. Finally, the NR decided to support both amplitude reporting scheme through higher layer configuration [16]:

- In wideband reporting mode, the UE does not report the differential amplitude on the subband (i.e., $p_{r,l,i}^{(SB)} = 1.$).
- For wideband + subband amplitude reporting, the UE reports differential amplitude on subband with 1bit quantization, namely $\left\{1, \sqrt{0.5}\right\}$, while wideband amplitude adopts 3-bit quantization and uniform quantizer with 3 dB step size to facilitate implementation, $\left\{1, \sqrt{0.5}, \sqrt{0.25}, \sqrt{0.125}, \sqrt{0.0625}, \sqrt{0.0313}, \sqrt{0.0156}, 0\right\}$.

The phase of each coefficient can be quantized by QPSK and/or 8PSK, which is determined by high-level configuration parameters. For wideband + subband amplitude reporting, the Type II codebook adopts nonuniform phase quantization [18]. High-precision (8PSK) phase quantization is used for strong beams while low-precision (QPSK) phase quantization is used on weak beams in order to optimize the trade-off between overhead and performance. If $L = \{2, 3, 4\}$, the first $K = \{4, 4, 6\}$ beams are quantized with 8PSK. The precoding vector only needs the differential phase and amplitude between different ports. Thus the dimension is $2L$-1. The Type II codebook takes the beam with the strongest wideband power as the reference "1", which is indicated by $\log_2(2L))$ bits. The UE reports the amplitude and phase of remaining $2L$-1 beam [17]. Similar to the UE selective report subband coding, the Type II codebook utilize a combination number to report L selected beams $\left(\log_2 C_{N1N2}^{L} \text{ bit}\right)$, which can reduce feedback overhead compared with the method bitmap indicator but at cost of certain complexity [19].

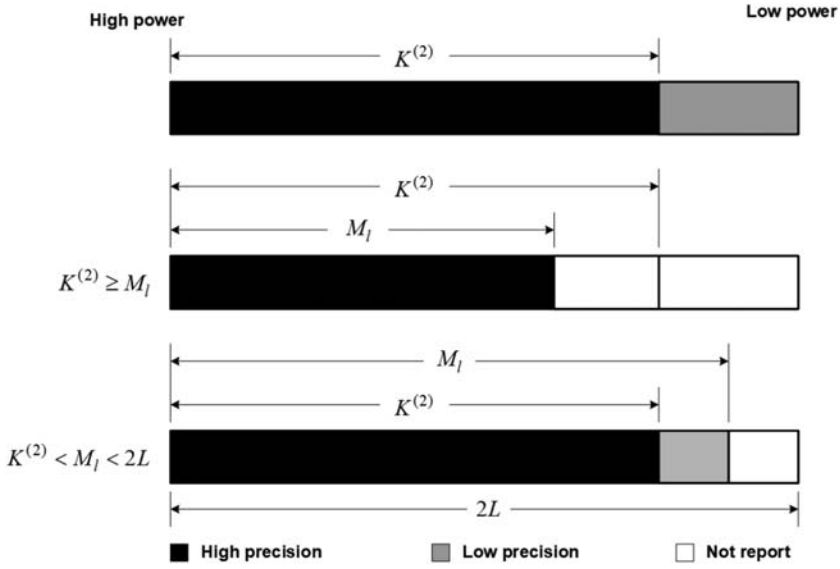


Figure 8.8 $\boldsymbol{W}_1$ Report the number of nonzero coefficients.

Because L is a high-layer parameter and the actual number of beams selected may be less than L, the wideband amplitude needs to include zero elements [20] to reduce the feedback overhead for the subband [21]. The wideband amplitudes in the two polarization directions are selected independently (allowed to be zero). In order for the network to determine the information length, the UE reports the number of wideband nonzero coefficients [22] (i.e., the number of L beams whose power is greater than zero). As shown in Fig. 8.8, M_l is the number of reported nonzero coefficients and $K^{(2)}$ is the number of 8PSK quantized beams.

8.1.3.3 Port selection codebook

If the network knows partial downlink channel information, for example, the direction of beam that is obtained through the UL/DL reciprocity, the UE can measure the beam-formed CSI−RS. In this case, the UE only needs to measure a few ports, and that can reduce complexity of calculating, storage, and also the CSI feedback overhead. On the basis of the Type II codebook, the $\boldsymbol{W}_1$ of port selection codebook is [23]:

$$\boldsymbol{W}_1 = \begin{bmatrix} \boldsymbol{E} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{E} \end{bmatrix}$$

The matrix $\boldsymbol{E}$ is expressed as:

$$\boldsymbol{E} = \left[\boldsymbol{e}_{\mathrm{mod}\left(md,\frac{X}{2}\right)}\ \boldsymbol{e}_{\mathrm{mod}\left(md+1,\frac{X}{2}\right)}\ \cdots\ \boldsymbol{e}_{\mathrm{mod}\left(md+L-1,\frac{X}{2}\right)}\right] \tag{8.5}$$

where X is the number of CSI−RS ports. In this method, the candidate ports are divided into d groups (d is high-layer configuration) for the UE to select continuous L ports. Each column of $\boldsymbol{E}$ has only one (i.e., e_i represents the position of port). In $\boldsymbol{E}$, L elements are one while the rest are zero, which is used to indicate the port selection.

8.1.3.4 CSI omission

One major issue of the Type II codebook is that the overhead is too high. As shown in Table 8.3, the feedback overhead of the Type II codebook is proportional to the rank. In some configurations, the payload is more than 500 bits, and the overhead of different configurations varies greatly. When the network side allocates PUSCH resources for CSI reporting, it is difficult to estimate the overhead accurately. Resources can only be allocated according to the maximum overhead, which will cause a waste of resources.

In order to solve this issue, NR considers some methods to reduce resource overhead. Since channel in frequency domain are correlated, the CSI of adjacent subbands are generally similar. Downsampling in subband can reduce overhead significantly. When the PUSCH resource allocation is insufficient, half of the subbands can be dropped, thereby reducing the overhead by about half, while ensuring there is no obvious distortion of the CSI. Specifically, CSI on the subband with even index are reported with high priority, while the CSI of odd subband may be omitted [24] (Fig. 8.9).

8.2 R16 codebook enhancement

Compared with the traditional codebook, the Type II codebook can significantly improve the accuracy of the channel information, thereby improving the performance of downlink transmission, especially for downlink multiuser transmission [25,26]. However, the feedback overhead of the Type II codebook is very high. For example, when $L = 4$, rank = 2 and the number of subbands is 10, the total feedback overhead reaches 584 bits. At the same time, the Type II codebook can only support single-layer and two-layer transmission due to high feedback cost,

Table 8.3 Feedback cost of Type II codebook.

Beam number	Over-sampling	Beam selection	Reference beam	Wideband amplitude	Wideband overhead	Subband amplitude	Subband phase	Total
Rank1 overhead (Number of bits)								
2	4	7	2	9	22	3	9	142
3	4	10	3	15	32	3	13	192
4	4	11	3	21	39	5	19	279
Rank2 overhead (Number of bits)								
2	4	7	4	18	33	6	18	273
3	4	10	6	30	50	6	26	370
4	4	11	6	42	63	10	38	543

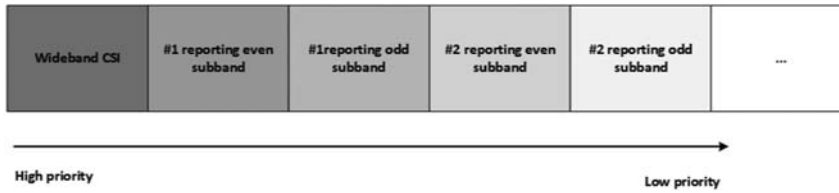


Figure 8.9 CSI priority.

which limits its application scenarios. In order to reduce the feedback signaling overhead of the Type II codebook and extend it to higher rank transmission, NR introduces an enhanced Type II codebook in R16, called the eType II codebook.

8.2.1 Overview of the eType II codebook

The Type II codebook compresses the SD of eigenvectors through L DFT beams, but does not compress the frequency-domain (FD) part. The overhead of the Type II codebook is mainly contributed by the coefficient matrix $\boldsymbol{W}2$, so the overhead of the Type II codebook can be reduced by compressing the linear combination coefficient of $\boldsymbol{W}2$ [27]. Therefore the eType II codebook considers two methods to reduce the overhead: firstly, compressing the FD complexity, and then selectively reporting some linear combination coefficients to compress the correlation of coefficient matrix.

In order to reduce the feedback overhead of the Type II codebook, two compression schemes were proposed at the RAN1#94bis meeting: the method of FD compression (FDC) and the method of time-domain compression (TDC). The key idea of both schemes is to reduce the feedback overhead by using FD correlation or time-domain sparsity. The FDC scheme uses the correlation of precoding coefficients between adjacent subbands and introduces a set of FD basis vectors to compress the FD complexity. The TDC scheme uses DFT or IDFT to convert the coefficients of the subbands into coefficients in the time-domain. If the DFT vectors are selected as the FD basis vectors, the FDC and TDC compression schemes are equivalent. Therefore at the RAN1#95 meeting, it was agreed to use the DFT vector as the basis vector for FDC matrix. Thus the FDC and TDC schemes were combined into one scheme [28–30], which was referred to as the FDC scheme.

At the RAN1#95 meeting, some companies also proposed another candidate compression scheme in order to reduce the FD feedback

overhead, which is a compression scheme based on singular value decomposition (SVD) [31]. This scheme uses SVD of $\boldsymbol{W}_2$ to obtain $\boldsymbol{W}_f$ and the compressed $\tilde{\boldsymbol{W}}_2$. Since this SVD scheme does not use a predefined set of basis vectors, the UE needs to dynamically report the basis vectors corresponding to $\boldsymbol{W}_f$ [32]. Although this scheme can capture the maximum signal energy and obtain the best compression effect, it is more sensitive to errors. Furthermore, the SVD scheme requires a relatively large feedback overhead for reporting the $\boldsymbol{W}_f$ of the subband. Compared with the aforementioned schemes, the SVD scheme has advantage when the feedback overhead is greater. Finally, the 3GPP adopted the FDC scheme as the FDC scheme.

According to the FDC scheme, the eType II codebook can be expressed as $W = \boldsymbol{W}_1 \tilde{\boldsymbol{W}}_2 \boldsymbol{W}_f^H$. The UE needs to report $\boldsymbol{W}_1 = \{b_i\}_{i=0}^{L-1}$, $\boldsymbol{W}_f = \{f_k\}_{k=0}^{M-1}$ and linear combination coefficients $\tilde{\boldsymbol{W}}_2 = \{c_{i,f}\}, i = 0, 1, \ldots, L-1, f = 0, 1, \ldots, M-1$, where

- The $\boldsymbol{W}$ has N rows and N_3 columns, where N is the number of CSI−RS ports, and N_3 is the number of FD units.
- $\boldsymbol{W}_1$ reuse the Type II codebook design and each polarization group contains L beams (i.e., $\boldsymbol{W}_1 = \begin{bmatrix} b_0 \ldots b_{L-1} & 0 \\ 0 & b_0 \ldots b_{L-1} \end{bmatrix}$.)
- Similar to the design of the Type II codebook, the $\tilde{\boldsymbol{W}}_2$ matrix contains all the $2L \times M$ coefficients of linear combination, where M is the number of FD basis vectors.
- $\boldsymbol{W}_f$ is the matrix of DFT basis vectors (i.e., $\boldsymbol{W}_f = \begin{bmatrix} f_0 \ldots f_{M-1} \end{bmatrix}$.)

For example, if rank = 1, the precoding matrix of the eType II codebook can be expressed as shown in Fig. 8.10 [33].

Fig. 8.11 shows the evolution from the Type II codebook to the eType II codebook. Compared with the Type II codebook, the eType II codebook adds a new part $\boldsymbol{W}_f$, where $\boldsymbol{H}$ is the eigenvector matrix, and the kth column of $\boldsymbol{H}$ is the channel eigenvector of the kth subband [27].

$$\underset{N \times N_3}{\mathbf{W}} = \underset{N \times 2L}{\mathbf{W}_1} \times \underset{2L \times M}{\tilde{\mathbf{W}}_2} \times \underset{M \times N_3}{\mathbf{W}_f}$$

Figure 8.10 General form of R16 eType II precoding matrix.

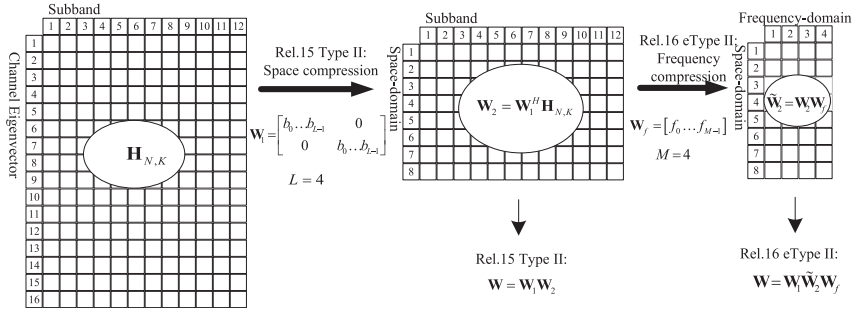


Figure 8.11 Schematic diagram of the evolution process from R15 to R16 Type II.

After determining the general form of the eType II codebook, the design of $\boldsymbol{W}_1$, $\tilde{\boldsymbol{W}}_2$, $\boldsymbol{W}_f^H$ will be discussed in next sections. The design of $\boldsymbol{W}_1$ is the same as that of the Type II codebook, so this section will not repeat the discussion on it. The design of $\tilde{\boldsymbol{W}}_2$ and $\boldsymbol{W}_f$ for rank = 1 will be first introduced, and then the case of rank > 1.

8.2.2 Frequency-domain matrix design

The critical part of $\boldsymbol{W}_f$ design is the selection of DFT basis vector (i.e., how to select M DFT basis vectors from N_3 FD units, where $N_3 = N_{sb} \times R$, $M = p \times \frac{N_3}{R}$, N_{sb} is the number of subbands, R indicates the number of FD units contained in each subband, and p is configured by high-level signaling to determine the number of DFT basis vectors).

For rank = 1, 3GPP considered the following two candidate schemes for the selection of DFT basis vectors:

- Common basis vector [34]: $2L$ beams select the same DFT basis vector, where $\boldsymbol{W}_f = [f_0 \ldots f_{M-1}]$, and M basis vectors are dynamically selected, as shown in Fig. 8.12.
- Independent basis vector [35]: each beam independently selects the DFT basis vector, $\boldsymbol{W}_f = [\boldsymbol{W}_f(0), \ldots, \boldsymbol{W}_f(2L-1)]$ where $\boldsymbol{W}_f(i) = [\boldsymbol{f}_{k_{i,0}} \boldsymbol{f}_{k_{i,1}} \ldots \boldsymbol{f}_{k_{i,M_i-1}}]$, $i \in \{0, 1, \ldots, 2L-1\}$, as shown in Fig. 8.13.

For the above two schemes, companies conducted a lot of simulations and found that the implementation of the scheme of common basis vector is simpler and the overhead is relatively small [36]. Therefore it was agreed to adopt the scheme of common basis vector for rank = 1 at the RAN1#AH1901 meeting.

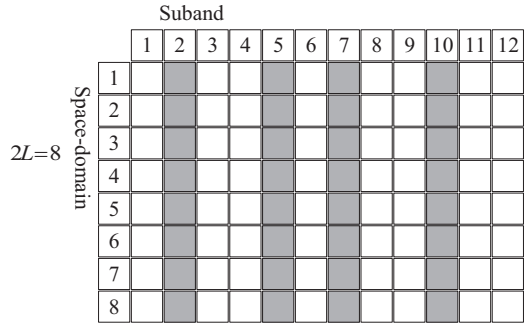


Figure 8.12 2*L* beams using the same DFT basis vector.

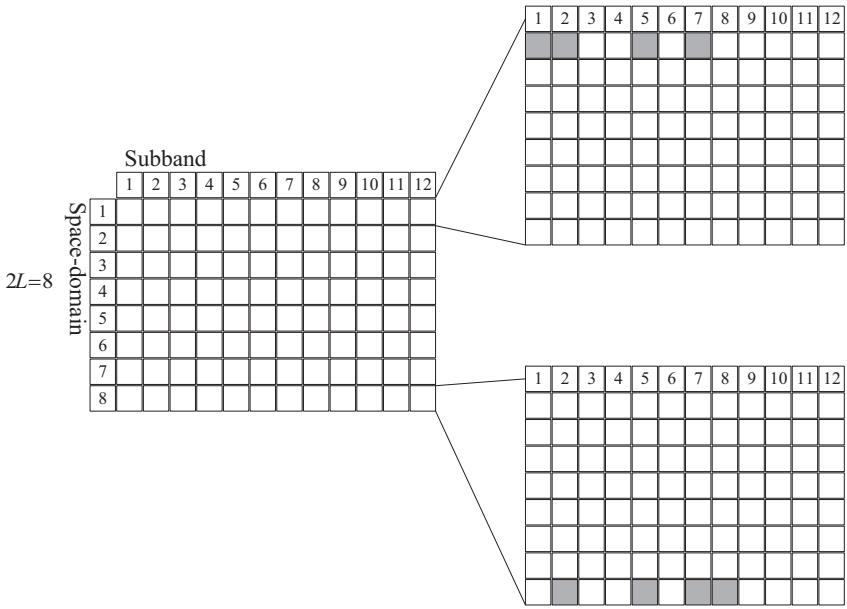


Figure 8.13 Each beam independently selects the DFT basis vector.

In addition, the companies also found that when selecting M DFT basis vectors in N_3 FD units, $\left\lceil \log_2 \binom{N_3 - 1}{M - 1} \right\rceil$ bits need to be fed back. If N_3 increases, the feedback bit will also increase. In order to further reduce the feedback overhead, a one-stage scheme and a two-stage window scheme that are suitable for different scenarios were agreed on at the RAN197 meeting [37]. When $N_3 \leq 19$, the one-step scheme is adopted,

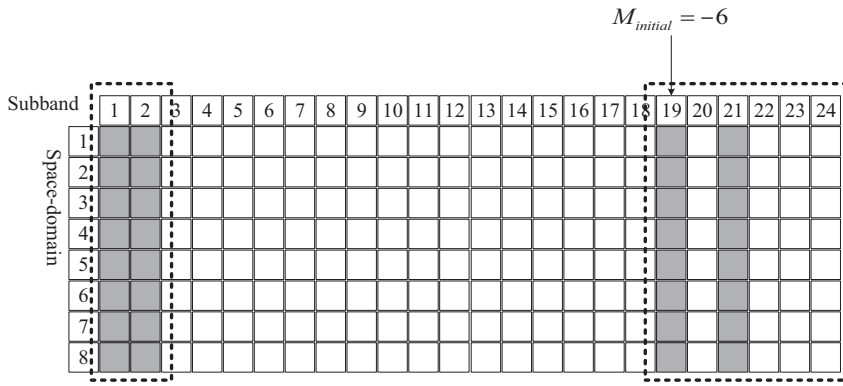


Figure 8.14 $N_3 > 19$, the two-stage window scheme.

as shown in Fig. 8.12. When $N_3 > 19$, the two-step windowing scheme is adopted. A window with a length of $2M$ is configured through high-level signaling, and the UE reports $M_{initial}$ to indicate the starting position of the window, and selectively reports M DFT basis vectors within that window, where $M_{initial} \in \{-2M+1, -2M+2, \ldots, 0\}$. An example is illustrated in Fig. 8.14. The black virtual frame is a window with a length of 2M, and the gray grid is the selected M DFT basis vectors.

8.2.3 Design of coefficient matrix

The W_2 matrix of the Type II codebook is a coefficient matrix containing $K = 2LM$ linear combinations, and each row contains related linear combination coefficients. Therefore a parameter K_0 is introduced in the eType II codebook in order to further reduce the overhead of the eType II codebook, where $K_0 = \beta \times 2LM$. A UE can select up to K_0 nonzero coefficients for reporting from a set of 2LM linear combination coefficients, where β is an RRC configuration parameter.

Regarding how to determine K_0, RAN1 discussed two schemes [38,39]:

- Unrestricted subset selection: K_0 nonzero coefficients are freely selected in the set of size $2LM$. An example is shown in Fig. 8.15 with $L = 4$, $M = 4$, $\beta = 0.25$.
- Polarization-common subset selection: the same nonzero coefficients are selected in both polarization directions. An example of this scheme is shown in Fig. 8.16.

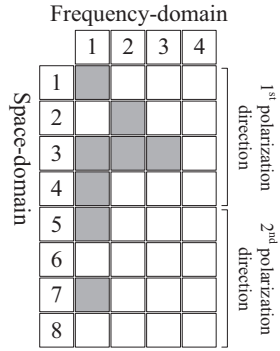


Figure 8.15 Unrestricted subset selection.

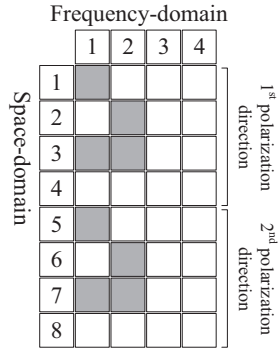
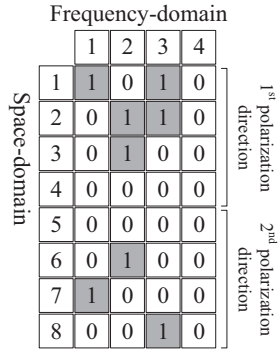


Figure 8.16 Polarization-common subset selection.

Based on the simulation evaluations, companies found that these two schemes do not have significant difference in terms of performance [40–42]. Although the scheme of common subset selection only needs a bitmap of size LM to report the exact location of zero or nonzero coefficients, it may happen that the coefficients in one polarization direction are nonzero while the same coefficients in the other polarization direction are zero. In order to avoid such a situation, at the RAN1#96 meeting, the method of unrestricted subset selection was adopted for K_0 selection. As shown in Fig. 8.17, the positions marked with 1 are the nonzero coefficients, while the positions marked with 0 are the zero coefficients.

After indicating the position of the nonzero coefficient, the UE needs to report the amplitude and phase of the corresponding nonzero coefficient. Three schemes were mainly considered in NR for the quantization of the nonzero coefficient. In the following discussion, the linear



	1	2	3	4
1	1	0	1	0
2	0	1	1	0
3	0	1	0	0
4	0	0	0	0
5	0	0	0	0
6	0	1	0	0
7	1	0	0	0
8	0	0	1	0

Figure 8.17 The 2*LM* bitmap indicates the location of nonzero coefficients.

combination coefficients related to beam $i \in \{0, 1, \ldots, L-1\}$ and FD units $f \in \{0, 1, \ldots, M-1\}$ are denoted as $c_{i,f}$, and the strongest coefficient is denoted as $c_{i*,f*}$.

The major features of quantization scheme 1 (which is similar to R15 Type II $\boldsymbol{W}_2$) are:

- The position of the strongest linear combination coefficient $(i*, f*)$ is indicated by $\log_2 K_{NZ}$ bits, where K_{NZ} is the actual number of nonzero coefficients reported. The strongest coefficient $c_{i*,f*} = 1$, so the UE does not need to report its amplitude and phase.
- For all the other linear combination coefficients $\{c_{i,f}, (i,f) \neq (i*, f*)\}$: the amplitude is quantized with three bits, and the phase is quantized with three bits (8PSK) or four bits (16PSK). The table of amplitude 3-bit quantization set is the same as that in R15.

The major features of quantization scheme 2 as shown in Fig. 8.18 [43] are:

- The position of the strongest linear combination coefficient $(i*, f*)$ is indicated by $\log_2 K_{NZ}$ bits. The strongest coefficient $c_{i*,f*} = 1$, so the UE does not need to report its amplitude and phase.
- Polarization-direction reference amplitude $p_{ref}(i,f)$: the reference amplitude of the polarization direction that contains the strongest coefficient is not reported. While in another polarization direction, the reference amplitude is quantized with four bits with respect to the strongest coefficient, and the quantization set table is $\left\{1, \left(\frac{1}{2}\right)^{\frac{1}{4}}, \left(\frac{1}{4}\right)^{\frac{1}{4}}, \left(\frac{1}{8}\right)^{\frac{1}{4}}, \ldots, \left(\frac{1}{2^{14}}\right)^{\frac{1}{4}}, 0\right\}$.
- For all the other coefficient $\{c_{i,f}, (i,f) \neq (i*, f*)\}$: in each polarization direction, the differential amplitude of the nonzero coefficient is

quantized with three bits with respect to the reference amplitude in that polarization direction. The quantization set table is $\left\{1, \frac{1}{\sqrt{2}}, \frac{1}{2}, \frac{1}{2\sqrt{2}}, \frac{1}{4}, \frac{1}{4\sqrt{2}}, \frac{1}{8}, \frac{1}{8\sqrt{2}}\right\}$. The final quantized amplitude is the multiplication of the reference amplitude and the differential amplitude: $p_{i,f} = p_{ref}(i,f) \times p_{diff}(i,f)$.

- The phase of each coefficient is quantized with three bits (8PSK) or four bits (16PSK).

The quantization scheme 3 [44] has the following major features:

- The position of the strongest linear combination coefficient $(i*, f*)$ is indicated by $\log_2 K_{NZ}$ bits, and the strongest coefficient $c_{i*,f*} = 1$, so the UE does not report its amplitude and phase.
- For $\left\{c_{i,f*}, i \neq i*\right\}$: the amplitude is quantized with four bits and the phase is quantized with 16PSK. The quantization set table is $\left\{1, \left(\frac{1}{2}\right)^{\frac{1}{4}}, \left(\frac{1}{4}\right)^{\frac{1}{4}}, \left(\frac{1}{8}\right)^{\frac{1}{4}}, \ldots, \left(\frac{1}{2^{14}}\right)^{\frac{1}{4}}, 0\right\}$.
- For $\left\{c_{i,f}, f \neq f*\right\}$: the amplitude is quantized with three bits and the phase is quantized with 8PSK or 16PSK. The quantization set table is $\left\{1, \frac{1}{\sqrt{2}}, \frac{1}{2}, \frac{1}{2\sqrt{2}}, \frac{1}{4}, \frac{1}{4\sqrt{2}}, \frac{1}{8}, \frac{1}{8\sqrt{2}}\right\}$.

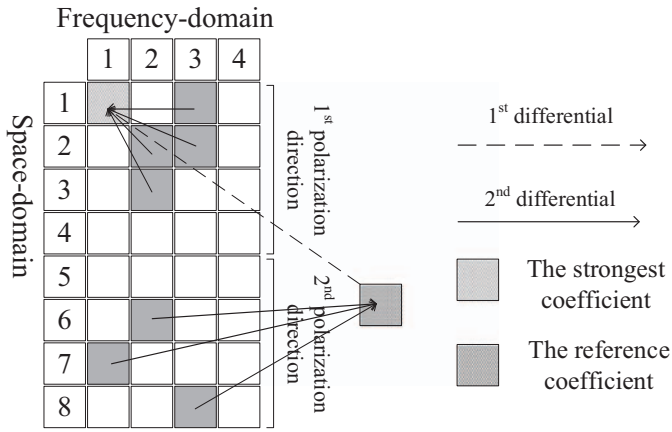


Figure 8.18 Quantization scheme 2.

Companies conducted a lot of simulation evaluations based on above schemes [40,45–47], and finally agreed on to adopt quantization scheme 2 at the RAN1#96 meeting.

According to the design in quantization scheme 2, the amplitude of the nonzero coefficient in $\tilde{\boldsymbol{W}}_2$ can be expressed as the multiplication of the reference amplitude and the differential amplitude. Assuming that the strongest coefficient is located in the first polarization direction (i.e., $c_{1,1}$ in the example shown in Fig. 8.1), the reference amplitude of its polarization direction is $p_0^{(1)} = 1$. The reference amplitude of the second polarization direction $p_1^{(1)}$ is quantized with four bits with respect to the strongest coefficient, as shown by the dotted line in the figure. If rank = 1, the specific expression form of $\tilde{\boldsymbol{W}}_2$ is:

$$\tilde{\boldsymbol{W}}_2 = \begin{bmatrix} \sum_{i=0}^{L-1}\sum_{f=0}^{M-1} c_{i,f} \\ \sum_{i=0}^{L-1}\sum_{f=0}^{M-1} c_{i+L,f} \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{L-1} p_0^{(1)} \sum_{f=0}^{M-1} p_{i,f}^{(2)} \varphi_{i,f} \\ \sum_{i=0}^{L-1} p_1^{(1)} \sum_{f=0}^{M-1} p_{i+L,f}^{(2)} \varphi_{i+L,f} \end{bmatrix}, \tag{8.6}$$

where $c_{i,f}$ is the linear combination coefficient, $p_0^{(1)}$ is the first polarization-direction reference amplitude, $p_1^{(1)}$ is the second polarization-direction reference amplitude, $p_{i,f}^{(2)}$ is the differential amplitude, and $\varphi_{i,f}$ is the phase of the nonzero coefficient.

8.2.4 Codebook design for rank = 2

For rank = 2, RAN1 further discussed the selection methods of SD subset, FD subset, and coefficient subset. The SD subset selection means selecting L spatial-domain DFT vectors in the beam set of N_1N_2, the FD subset selection means selecting M DFT basis vectors in N_3 FD units, and coefficient subset selection means selecting K_{NZ} nonzero coefficients from the *2LM* linear combination coefficient set. Although some simulation results show that it can improve the performance by selecting SD subsets independently in each layer, R16 still uses the same beam in different layers, which are consistent with R15 due to the overhead and complexity. Three different schemes for selecting FD subset and coefficient subset were considered:

- The selection of FD subset and coefficient subset is common to all the layers.
- FD subset selection is common to all the layers, but the coefficient subset selection is independent of each layer.
- The selection of FD subset and coefficient subset is independent for each layer.

For rank = 2, the independent selection of FD subsets and coefficient subsets for each layer can provide sufficient configuration flexibility, and can provide considerable performance gain without increasing the feedback overhead by too much. Considering the tradeoff between performance and overhead, the independent selection of the FD subset and coefficient subset scheme of each layer was adopted at the RAN1#96 meeting [48–50]. K_0 of the two layers is the same, where $K_0 = \beta \times 2LM$ (β is configured by RRC signaling), and the number of nonzero coefficients reported of each layer cannot exceed K_0, (i.e., $K_l^{NZ} \leq K_0$, $l = 1, 2$).

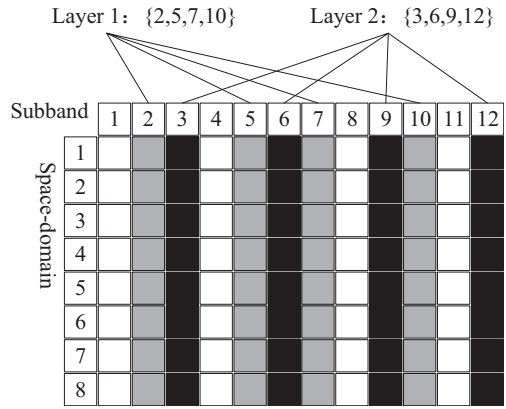


Figure 8.19 Independent FD subset selection (rank = 2).

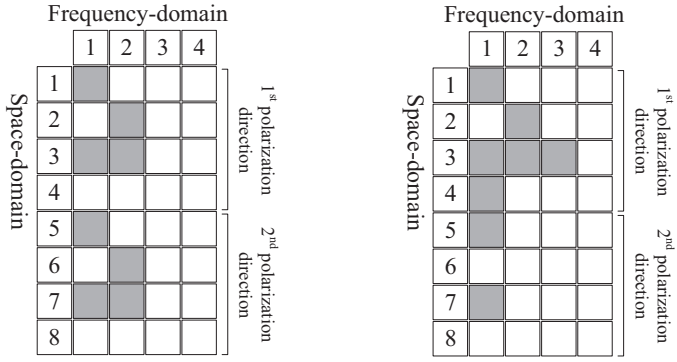


Figure 8.20 Independent coefficient subset selection (rank = 2).

8.2.5 Codebook design for high rank

The low-rank eType II codebook design was introduced is previous sections, and can significantly reduce the overhead compared to the Type II codebook. In order to provide better performance in scenarios with good channel quality, the eType II codebook introduces enhancements for high rank and extends the design for rank = 1/2 to rank = 3/4. The method of SD subset selection, FD subset selection, and coefficient subset selection of high rank follow the same principle as for rank = 2. SD subset selection is the same for each layer, while FD subset selection and coefficient subset selection are independent in each layer.

Since the overhead of the codebook is proportional to the number of nonzero coefficients and quantization parameters, so directly extending the design of rank = 1/2 to rank = 3/4 will bring a significant increase in the overhead. Considering that the purpose of introducing the eType II codebook in R16 is to reduce the overhead of the Type II codebook, it was decided at the RAN1#96bis meeting that extending the design to high rank will not increase the feedback overhead. Specifically, the actual number of nonzero coefficients reported for each layer will not be greater than K_0, (i.e., $K_l^{NZ} \leq K_0$,) and the total number of nonzero coefficients reported for all layers cannot be greater than $2K_0$, (i.e., $\sum_{l=1}^{R} K_l^{NZ} \leq 2K_0$) [51,52]. In order to meet the above two restrictions, after considering the complexity and performance, the final determined parameter combination (L, P, β) is shown in Table 8.4 [53].

Table 8.4 Parameter combination of eType II codebook.

***paramCombination* − *r*16**	***L***	p_v		β
		$v \in \{1, 2\}$	$v \in \{3, 4\}$	
1	2	1/4	1/8	1/4
2	2	1/4	1/8	1/2
3	4	1/4	1/8	1/4
4	4	1/4	1/8	1/2
5	4	1/4	1/4	3/4
6	4	1/2	1/4	1/2
7	6	1/4	–	1/2
8	6	1/4	–	3/4

Table 8.5 Expression of eType II codebook.

Layer	W
$v = 1$	$W^{(1)} = W^1$
$v = 2$	$W^{(2)} = \frac{1}{\sqrt{2}}\left[W^1 W^2\right]$
$v = 3$	$W^{(3)} = \frac{1}{\sqrt{3}}\left[W^1 W^2 W^3\right]$
$v = 4$	$W^{(4)} = \frac{1}{\sqrt{4}}\left[W^1 W^2 W^3 W^4\right]$

8.2.6 eType II codebook expression

The eType II codebook can be expressed by Table 8.5.where $W^l = W_1 \tilde{W}_2^l \left(W_f^l\right)^H = \frac{1}{\sqrt{N_1 N_2 \gamma_{t,l}}} \begin{bmatrix} \sum_{i=0}^{L-1} v_{m_1^{(i)},m_2^{(i)}} p_{l,0}^{(1)} \sum_{f=0}^{M_v-1} \gamma_{t,l}^{(f)} p_{l,i,f}^{(2)} \varphi_{l,i,f} \\ \sum_{i=0}^{L-1} v_{m_1^{(i)},m_2^{(i)}} p_{l,1}^{(1)} \sum_{f=0}^{M_v-1} \gamma_{t,l}^{(f)} p_{l,i+L,f}^{(2)} \varphi_{l,i+L,f} \end{bmatrix}$, $l = 1, \ldots, v$, v is the layer number, $i = 0, 1, \ldots, L-1, f = 0, 1, \ldots, M_v - 1$, $t = 0, \ldots, N_3 - 1$, $\gamma_{t,l} = \sum_{i=0}^{2L-1} \left(p_{l,\frac{i}{L}}^{(1)}\right)^2 \left| \sum_{f=0}^{M_v-1} \gamma_{t,l}^{(f)} p_{l,i,f}^{(2)} \varphi_{l,i,f} \right|^2$. $v_{m_1^{(i)},m_2^{(i)}}$ is the 2D-DFT beam. $\gamma_{t,l} = \left[\gamma_{t,l}^{(0)} \gamma_{t,l}^{(1)} \ldots \gamma_{t,l}^{(M_v-1)}\right]$ is the M_v DFT basis vectors of the layer l. $p_{l,0}^{(1)}$ is the reference amplitude of the layer l in the first polarization direction and $p_{l,1}^{(1)}$ is the reference amplitude of the layer l in the second polarization direction. $p_{l,i,f}^{(2)}$ is the differential amplitude of the layer l, $\varphi_{l,i,f}$ is the phase of the nonzero coefficient of layer l. For all the zero coefficients, $p_{l,i,f}^{(2)} = 0$, $\varphi_{l,i,f} = 0$.

8.3 Beam management

With the rapid development of mobile communication technology and the wide deployment of mobile network, radio spectrum resources are becoming increasingly scarce since the low-frequency band with low pathloss and good coverage is almost used up. In order to offer higher transmission rate and larger system capacity, the new generation of mobile communication tends to exploit middle-frequency band and high-frequency band, such as the middle-frequency band of 3.5−6 GHz and

the millimeter wave band. The middle and high-frequency bands have the following major characteristics (Fig. 8.18):

- Due to the relatively rich frequency resource in middle-frequency and high-frequency bands, it is easy to allocate continuous frequency resource with larger bandwidth, which is beneficial for the commercial deployment and UE experience.
- Due to higher frequency, the larger pathloss will lead to a smaller coverage. Therefore better isolation in the space dimension can be achieved and it can facilitate the deployment of ultradense network.
- Due to the high-frequency band, the corresponding antenna and other hardware modules are small. Thus a larger number of antennas can be used for NR, which is conducive to the realization of the massive MIMO technology (Fig. 8.19).

As can be seen from the characteristics described above, the high frequency leads to the problem of limited coverage. Furthermore, the large number of antennas brings the problem of huge complexity and cost of implementation. In order to effectively solve these two problems, Analog Beamforming technology is adopted in NR, which can enhance network coverage and reduce the implementation complexity of devices (Fig. 8.20).

8.3.1 Overview of analog beam-forming

The basic principle of analog beam-forming technology is to adjust the phase/amplitude of the output of each antenna element (e.g., through the phase shifter), so that a group of antennas can form the beam in different directions and improve cell coverage through beam sweeping. By using phase shifter to form beams in different directions, the high complexity of traditional MIMO schemes when applying in a wide bandwidth can be avoided, where the traditional MIMO schemes can also be referred to as digital beam-forming.

An example of analog beam-forming is shown as Fig. 8.21. The left is a traditional LTE or an NR system without analog beam-forming, and the right is an NR system using analog beam-forming:

- As shown in the left of Fig. 8.21, the LTE/NR network uses a wide beam to cover the entire cell. Users 1−5 can receive the network signal at any time.
- As shown in the right of Fig. 8.21, the network uses narrow beams (such as beams 1−4 in the figure), and uses different beams to cover different areas within the cell at different time. For example, at time 1,