3GPP TSG RAN WG1 #118 R1-2405975

Maastricht, Netherlands, August 19th–August 23th, 2024

Source: CMCC

Title: Discussion on specification support for beam management

Agenda item: 9.1.1

Document for: Discussion & Decision

# Introduction

In RAN1#117 meeting [1], the following agreements were made regarding the spec impacts of AI based beam management.

|  |
| --- |
| Agreement  For BM-Case1 and BM-Case2 with a UE-side AI/ML model:   * Support Type 1 performance monitoring, including the following two options:   + Option 1 (NW-side performance monitoring):     - UE sends a report to NW (for the calculation of performance metric at NW)       * Measurement results from resource set for monitoring, e.g., L1-RSRP and/or RS index is supported as the content of the report       * FFS on other contents     - The report is at least configured/triggered by NW     - Note: this may or may not have additional spec impact   + Option 2 (UE-assisted performance monitoring):     - UE calculates performance metric(s)       * FFS how to report and what to report   + FFS whether to trigger the report based on event(s) for Option 1 and/or Option 2 * FFS Type 2 performance monitoring   Agreement  At least for NW sided model, for the quantization of a reported L1-RSRP value at least for the report in L1 signaling, support   * Support differential L1-RSRP reporting with legacy quantization step and range   + FFS: larger quantization step(s) than the already supported legacy quantization step for differential L1-RSRP and/or for absolute L1-RSRP   + FFS: Smaller range(s) for differential L1-RSRP than the already supported legacy range   Agreement  Following Working Assumption is confirmed.  Working Assumption  For report content of inference results for UE-sided model for BM-Case 2, the RSRP ofpredicted beam(s) in the report of inference results, is the predicted RSRP, where the predicted RSRP is based on AI/ML output.  Agreement  For NW-sided model, for inference report, at least for BM-Case 1, the content in a beam report in L1 signaling, support   * L1-RSRPs and corresponding beam information of Top M beam(s) with largest M measured value(s) of L1-RSRP(s) of a measurement resource set, where M is configured by gNB   + If M = the size of the measurement resource set, the content is all L1-RSRPs and one beam index (i.e., CRI/SSBRI) for the largest measured value of L1-RSRP of a measurement resource set * FFS: L1-RSRPs and corresponding beam information of up to M beams within X dB gap to the largest measured value of L1-RSRP, X and M are configured by gNB, and whether/how to report number of reported beams * FFS on the maximum value of M (where M can be larger than 4) based on UE capability (M may or may not be different for different reporting contents) * FFS on beam information * Note: Purpose, such as above “For NW-sided model, for inference report, at least for BM-Case 1”, will not be specified in RAN 1 specifications |

In this contribution, we concentrate on discussing the procedure and the potential specification impact of AI based beam management during different LCM operation.

# Discussion on specification impact of data collection

In this section, we discuss the specification impact of data collection for training, inference and monitoring.

## Data collection of NW-side model for training

### 2.1.1 Reporting Content

In RAN1#117 meeting, FL has provided following proposal about content for NW-side data collection for training for higher layer report.

|  |
| --- |
| Proposal 3.4c  For content for data collection for NW-sided model, at least for BM-Case 1, further study the following options:   * Opt 1: L1-RSRPs from RS resources from one or two sets of beams configured to UE   + FFS on whether/how the corresponding beam information needs to be reported explicitly or implicitly.   + Note: overhead reduction is not precluded. * Opt 2: L1-RSRPs from RS resources from one set of beams configured to UE, and beam information of Top K from RS resources from another set of beams   + K is configurable. FFS: K values   + FFS on whether/how the corresponding beam information to L1-RSRPs need to be reported explicitly or implicitly. * Note: the max number L1-RSRPs measurements and reporting |

Data collection content for different purposes can be different. For example, data collection content for inference is only related to L1-RSRPs and/or beam information based on one resource set, data collection content for monitoring depends on monitoring KPI. We prefer to discuss data collection content for different purposes independently, and focus on data collection content for training in this section.

With opt 1, RS of set A and set B can be configured in one measurement resource set regardless set A and set B are different or set B is subset of set A. We do not need to distinguish set A and set B in two resource set. The time interval between training and inference/monitoring can be long, if inference and monitoring need independent set A and set B, gNB can configure set A and set B separately afterwards.

One issue needs clarification is whether L1-RSRPs of all RS are reported or a subset of L1-RSRPs and corresponding beam information are reported considering reporting overhead reduction. For higher layer report, payload size of UL resource is sufficient, L1-RSRPs of all RS can be reported. With legacy differential quantization method, beam information is CRI of RS with largest L1-RSRP.

For L1 signaling, since existing CSI framework can be reused and latency of report is small, L1 signaling is also supported for training data collection. To reduce reporting overhead of each sample/instance, reported number of beams can be reduced by setting RSRP threshold (low RSRP that contributes less to beam prediction is omitted) or new quantization method (e.g. temporal difference among multiple past time instances for BM-Case2). Bitmap can be used as beam information to represent RSRP of which beam is reserved.

Another issue needs clarification is current wording “one set of beam” does not exclude to configure multiple sets to constitute set A. As discussed in section 2.1.2, configuration of set A may need enhancement (e.g. more than one measurement resource set) if beam number of set A can be larger than 64.

With opt 2, the intention is to report L1-RSRPs of set B and Top K RS ID of set A.Set A and set B are configured separately. The value of K can be configured by RRC and is typically 2~10 during SI. For higher layer report, reporting overhead reduction is not needed, beam information is CRI of RS with largest L1-RSRP in measurement set and CRI of Top K RS in prediction set. For L1 signaling, considering one assumption is Set B is a subset of measured beam Set C, a subset of L1-RSRPs and corresponding beam information may be reported. Bitmap can be used as beam information to represent RSRP of which beam is reserved in measurement resource set.

Above all, depending on different output of AI model, option 1 and 2 are supported, where multiple sets constituting one beam set (i.e. set A) is not excluded. If L1 signaling is used for data collection, reporting overhead reduction is supported.

**Proposal 1: L1 signaling is supported for NW-sided training data collection.**

**Proposal 2: For NW-sided model, for data collection for training, at least for BM-Case1, option 1 and 2 are supported for the report contents:**

* **Opt 1: L1-RSRPs from RS resources from one set of beams configured to UE**
  + **for higher layer report, beam information is CRI of RS with largest L1-RSRP**
  + **for L1 signaling, beam information can be bitmap**
* **Opt 2: L1-RSRPs from RS resources from one set of beams configured to UE, and beam information of Top K from RS resources from another set of beams**
  + **K is configured by RRC, value of K is 1~10**
  + **for higher layer report, beam information is CRI of RS with largest L1-RSRP in measurement set and CRI of Top K RS in prediction set**
  + **for L1 signaling, beam information is bitmap and CRI of Top K RS in prediction set**

**Proposal 3: If L1 signaling is applied for NW-sided training data collection, further discuss reporting all or a subset of L1-RSRPs from the resource set, at least including data omission, e.g., L1-RSRP(s) higher than a threshold.**

### 2.1.2 Configuration of Set A

No matter best or quasi-optimal Rx beam assumption is considered, beam number of set A may be larger than 64. Existing UE capability for RS measurement and CSI reporting framework need enhancement.

As described in TS 38.213 and TS 38.331, existing UE capability on maximum number of RS for RSRP measurement is up to 64 for each resource set and up to 64 for all resource set across all CCs. With quasi-optimal Rx beam assumption, UE capability for RS number per resource set needs extension and only one periodic/semi-persistent resource set is needed, or the number of resource set in one *CSI-ResourceConfig* needs extension, to support measurement of set A with more than 64 RS.

|  |
| --- |
| **TS 38.213** **5.2.1.4.1 Resource Setting configuration** A UE is not expected to be configured with more than 64 NZP CSI-RS resources and/or SS/PBCH block resources in resource setting for channel measurement for a CSI-ReportConfig with the higher layer parameter *reportQuantity* set to 'none', 'cri-RI-CQI', 'cri-RSRP', 'ssb-Index-RSRP', 'cri-SINR' or 'ssb-Index-SINR'.  If the UE is configured with a *CSI-ReportConfig* with the higher layer parameter *reportQuantity* set to 'cri-RSRP' or 'ssb-Index-RSRP',  - if the UE is configured with the higher layer parameter *groupBasedBeamReporting* set to 'disabled', the UE is not required to update measurements for more than 64 CSI-RS and/or SSB resources, and the UE shall report in a single report *nrofReportedRS* (higher layer configured) different CRI or SSBRI for each report setting.  - if the UE is configured with the higher layer parameter *groupBasedBeamReporting* set to 'enabled', the UE is not required to update measurements for more than 64 CSI-RS and/or SSB resources, and the UE shall report in a single reporting instance two different CRI or SSBRI for each report setting, where CSI-RS and/or SSB resources can be received simultaneously by the UE either with a single spatial domain receive filter, or with multiple simultaneous spatial domain receive filters. **5.2.1.4.3 L1-RSRP Reporting** For L1-RSRP computation  - the UE may be configured with CSI-RS resources, SS/PBCH Block resources or both CSI-RS and SS/PBCH block resources, when resource-wise quasi co-located with 'QCL-Type C' and 'QCL-TypeD' when applicable.  - the UE may be configured with CSI-RS resource setting up to 16 CSI-RS resource sets having up to 64 resources within each set. The total number of different CSI-RS resources over all resource sets is no more than 128.  **TS 38.331**  BeamManagementSSB-CSI-RS ::= SEQUENCE {  maxNumberSSB-CSI-RS-ResourceOneTx ENUMERATED {n0, n8, n16, n32, n64},  maxNumberCSI-RS-Resource ENUMERATED {n0, n4, n8, n16, n32, n64},  maxNumberCSI-RS-ResourceTwoTx ENUMERATED {n0, n4, n8, n16, n32, n64},  supportedCSI-RS-Density ENUMERATED {one, three, oneAndThree} OPTIONAL,  maxNumberAperiodicCSI-RS-Resource ENUMERATED {n0, n1, n4, n8, n16, n32, n64}  } |

With best Rx beam assumption, existing periodic/semi-persistent CSI reporting framework can not support unified Tx beam and Rx beam sweeping with more than 64 RS within a short time. Based on existing spec, one periodic/semi-persistent *CSI-ReportConfig* is associated with one resource set. If repetition is on, the resource set is configured for Rx beam sweeping. If repetition is off, the resource set is configured for Tx beam sweeping, the Rx beam determination is up to UE implementation. If multiple *CSI-ReportConfig* associated with resource set with repetition on is used for unified beam sweeping, the required number of *CSI-ReportConfig* is as large as the number of Tx beams. If multiple *CSI-ReportConfig* associated with resource set with repetition off is used for unified beam sweeping, the required number of *CSI-ReportConfig* is small, e.g. the same as the number of Rx beams. But UE does not assume to switch Rx beam among different resource set, resulting in incomplete unified beam sweeping. Thus, CSI reporting enhancement is needed to support unified beam measurement for data collection.

With quasi-optimal Rx beam assumption, only Tx beam sweeping is needed, Rx beam can refer to QCL source. To perform Tx beam sweeping with more than 64 RS within each period, one possible way is to extend maximum number of RS for RSRP measurement per resource set. Time offset of each RS can be different so that the number of RS in each slot still satisfies measurement of maximum 64 RS in each slot. Another possible way is to configure multiple resource sets in one *CSI-ResourceConfig* to constitute set A, where time offset of each resource set is different. This enhancement on CSI reporting framework is also suitable for configuration of set A for inference and monitoring.

**Proposal 4: Regarding data collection for NW-side model, following options can be considered for the configuration of Set A:**

* **Option 1: The UE capability of the maximum number of the RS per resource set for the RSRP measurement can be enhanced.**
* **Option 2: Multiple resource sets can be supported in one *CSI-ResourceConfig***

### 2.1.3 Configuration of Set B

With NW side model, the set B is cell-specific and determined by gNB. During the SI phase, both fixed set B pattern and pre-configured multiple Set B patterns were studied. The pre-configured multiple Set B patterns were also proved with high beam prediction accuracy. Besides, one AI model with good generalization capability may support multiple set B with different beam number as input. Then, the configuration of multiple Set B or multiple Set B patterns can be supported.

Reusing existing CSI reporting framework, one method is gNB configures one resource set including RS of all set B patterns and configures different set B patterns with bitmap of the resource set. gNB may indicate UE to measure set B patterns in pre-defined order (e.g. in ascending order of beam pattern ID) in each time instance, or UE determines to measure one set B pattern and reports beam pattern ID together with L1-RSRPs.

**Proposal 5: Regarding training data collection for NW-side model, the configuration of multiple set B or multiple Set B patterns can be supported.**

### 2.1.4 Rx beam assumption

Regarding Rx beam assumption for NW-side model, due to evaluation results of R18 SI phase, best or quasi-optimal Rx beam has good beam prediction accuracy. Quasi-optimal Rx beam can save RS overhead without sweeping all Rx beams and is similar as existing mechanism where Rx beam is determined by UE implementation. Best Rx beam needs larger RS overhead for unified Tx and Rx beam sweeping, but has benefit in generalization of Rx antenna configuration. If gNB wants to have flexibility to configure data collection with best Rx beam, configuration of set A needs further design as discussed in section 2.1.2. If gNB intends to save RS overhead and use quasi-optimal Rx beam assumption, an indication may be needed in CSI-ReportConfig to align the Rx beam assumption between gNB and UE, in case UE determines Rx beam by implementation and results in poor performance of AI model. Above all, Rx beam assumption for a measurement report can be up to gNB implementation.

**Proposal 6: Rx beam assumption for a measurement report can be up to gNB implementation. If quasi-optimal Rx beam assumption applies, an indication may be needed in CSI-ReportConfig to align the Rx beam assumption between gNB and UE.**

## Data collection of UE-side model for training

### 2.2.1 UE request for data collection

If data collection is performed at the UE side, there is no need to disclose the detailed Rx beam information to the network, but some basic information like the required number of samples, the required number of Tx beams in Set B and Set A (included in meta info) should be reported to the network.

Different AI models/algorithms may have different preferred set B pattern, and UE reports model-specific preferred set B pattern when requests for data collection for training. If set B pattern transmitted by gNB is determined by AI model/algorithm, the RS overhead during training and inference is large. To reduce RS overhead at gNB, gNB can pre-configure several set B patterns, and then choose to transmit an appropriate set B pattern to UE for generating the input data of AI model.

**Proposal 7: Regarding data collection for UE-side model for training, UE can request for preferred set B.**

**Proposal 8: Regarding data collection for UE-side model for training, set B as AI model input is determined by gNB.**

### 2.2.2 configuration of set A and set B

Similar as data collection for NW-side model, measurement of Tx beams in a short time for each sample is preferred, UE capability on maximum number of RS for RSRP measurement per resource set or the number of resource set needs enhancement, the configuration method of set A for NW-side model can be reused.

**Proposal 9: Regarding data collection for UE-side model, the configuration method of set A for NW-side model can be reused.**

* **Option 1: The UE capability of the maximum number of the RS per resource set for the RSRP measurement can be enhanced.**
* **Option 2: Multiple resource sets can be supported in one *CSI-ResourceConfig***

There are four options to configure the association of beams within set B and beams within set A.

* + Option 1: one *CSI-ResourceConfigId* is configured for both Set A and Set B
  + Option 2: two *CSI-ResourceConfigId* s are configured for Set A and Set B separately
  + Option 3: one *CSI-ResourceConfigId* is configured for Set B, Set A is configured using separate resource set(s) other than that represented by *CSI-ResourceConfigId*
  + Option 4: separate *CSI-ReportConfig* for Set A and Set B

With option 1, configuring separate resource sets for Set A and Set B for measurement and prediction purposes within one *CSI-ResourceConfig* is supported, which is similar as the existing CSI resource setting with one set for channel measurement and one set for inference measurement. Comparing with joint resource set for Set A and Set B, UE can easily distinct resource for measurement and resource for prediction .

There can be three methods to configure the association of beams within set B and beams within set A. Method 1 is indicating association of set B and set A with TCI state. This is suitable for the case set B is different from set A, according to the definition of QCL source, set B (e.g. SSB) can naturally be used as QCL source of set A. For the case set B is subset of set A, configuring the association of set A/B beams based on TCI state is a little confusing, since TCI state in this case means one RS and its QCL source are the same RS, which is different from the definition of QCL source. Method 2 is indicating association of set B and set A with RS ID, e.g. indicating each RS in set B with attached RS ID within set A, or indicating set B with RS ID subset of set A. Method 2 is suitable for the case set B is subset of set A and the case set B is different from set A. Method 3 is indicating set B with a bitmap of set A, which adapts to the case that set B is subset of set A. To represent the association more explicitly, association with RS ID or bitmap is preferred than with TCI state.

With option 2, separate resource sets are configured in two *CSI-ResourceConfig* for Set A and Set B separately. The association of set B and set A is restricted to RS ID. Bitmap is not suitable since UE does not know the bitmap in second *CSI-ResourceConfig* corresponds to which set in the first *CSI-ResourceConfig.*

Option 3 is similar as option 2, separate resource sets are configured for Set A and Set B. Differently, resource set for Set A is not within second *CSI-ResourceConfig.* During data collection for training, set A is the set for prediction, but set A also needs measurement to obtain practical RSRP, thus configuring Set A outside *CSI-ResourceConfig* is confusing. The motivation and benefit of option 3 is not clear.

Option 4 is deprioritized since for UE sided model, measurement results of set A and set B do not need report. One *CSI-ReportConfig* is enough to configure measurement of set A and set B, there is no reason to configure separate *CSI-ReportConfig* for set A and set B.

**Proposal 10: For UE-sided model at least for BM Case-1, following two options are supported for the configuration of training data collection.**

* + **Option 1: one *CSI-ResourceConfigId* is configured for both Set A and Set B**
    - **separate resource sets are configured for Set A and Set B in *CSI-ResourceConfig***
  + **Option 2: two *CSI-ResourceConfigId* s are configured for Set A and Set B separately**

**Proposal 11: Regarding to option 1 of training data collection for UE-sided model, indication of association of Set A and Set B can be based on RS ID or bitmap.**

**Proposal 12: Regarding to option 2 of training data collection for UE-sided model, indication of association of Set A and Set B can be based on RS ID.**

# Discussion on specification impact of inference

Model inference may perform at NW side or UE side. To increase the prediction accuracy, Top-K best beams can be predicted by AI model and the best beam can be selected by measuring the L1-RSRP of Top-K best beams. In this section, we discuss the necessity of Top-K beam sweeping procedure during inference and specification impact of inference.

## Model inference at NW side

The model inference procedure of Tx beam prediction at NW side is shown in Fig. 1.

* Firstly, gNB transmits sparse Tx beams in Set B, where Set B beam pattern is cell-specific.
* In step 2, UE reports the L1-RSRP of measured beams.
* In step 3, gNB inputs L1-RSRP of measured beams into AI model and outputs the index and L1-RSRP of Top K best beams of Set A..
* In step 4, gNB jointly sweeps Top K beams and all Rx beams. UE measures L1-RSRP of Top K best beams.
* In step 5, UE reports the index and L1-RSRP of practical Top 1 best beam or part of Top K best beams. Legacy beam reporting mechanism is reused.
* In step 6, gNB indicates the beam for DL data transmission. Legacy beam indication mechanism is reused.

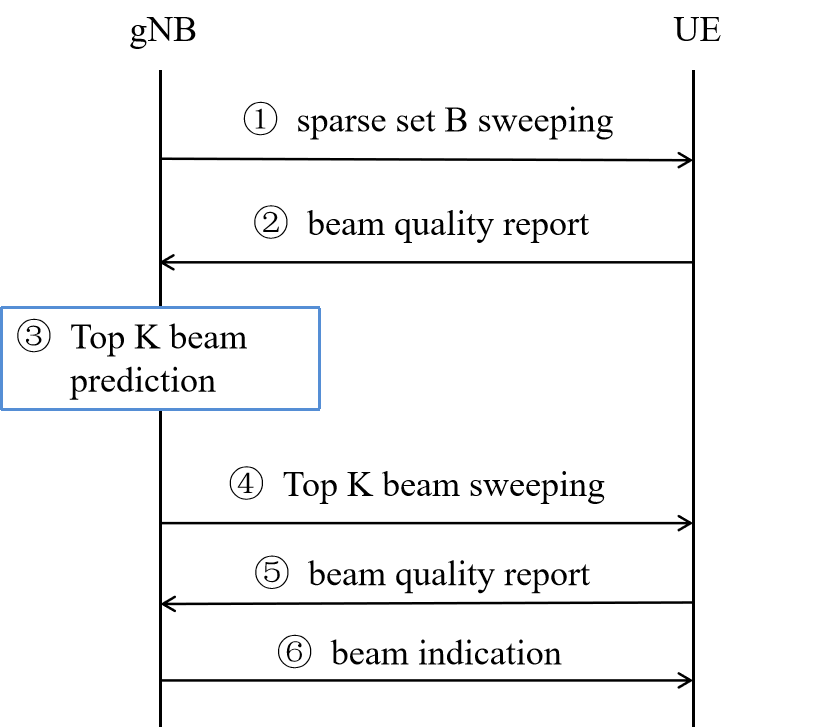


Fig. 1 inference at NW side

### 3.1.1 necessity of Top-K beam sweeping procedure

One controversial issue is whether top K beam sweeping is necessary. The comparison of with and without top K beam sweeping procedure from the perspective of beam prediction accuracy, RS overhead, delay, TCI state indication is summarized in Table 1.

Table 1 comparison of with and without top K beam sweeping procedure

|  |  |  |
| --- | --- | --- |
|  | with top K beam sweeping | without top K beam sweeping |
| beam prediction accuracy | * >90% | * ~80% |
| RS overhead | * comparatively large additional RS measurement overhead | * small RS measurement overhead |
| delay | * Top K beam sweeping delay * known TCI state switching delay is short | * unknown TCI switching delay includes RSRP measurement and SSB measurement delay as defined in 38.133 * The delay difference is small between with and without top K beam sweeping |
| TCI state indication | * if target TCI state is in the active TCI state list for PDSCH/PDCCH: DCI based beam indication * if target TCI state is not in the active TCI state list for PDSCH/PDCCH: MAC-CE based beam activation | * known target TCI indication: DCI or MAC-CE * unknown target TCI indication: MAC-CE based beam activation |

With top K beam sweeping, gNB can indicate top 1 genie-aided beam based on measured RSRP. Top K/1 beam prediction accuracy can be larger than 90% if K is approximately chosen. Top K beam sweeping introduces comparatively large additional RS measurement overhead for UE. Considering the worst case, gNB needs to transmit aperiodic RS in set A after each round of inference.

Between inference and new target TCI state applies, Top K beam sweeping delay and known TCI state switching delay are included. As defined in 38.133, if the target TCI state is known, upon receiving PDSCH carrying MAC-CE activation command in slot n, UE shall be able to receive UE-dedicated PDCCH/PDSCH with target TCI state of the serving cell on which TCI state switch occurs at the first slot that is after slot n+ THARQ + + TOk\*(Tfirst-SSB + TSSB-proc) / *NR slot length,* where

- Tfirst-SSB is time to first SSB transmission after MAC CE command is decoded by the UE; The SSB shall be the QCL-TypeA or QCL-TypeC to target TCI state

- TSSB-proc = 2 ms;

- TOk = 1 if target TCI state is not in the active TCI state list for PDSCH/PDCCH, 0 otherwise.

If target TCI state is in the active TCI state list for PDSCH/PDCCH, TCI state switch delay is short and DCI based beam indication is performed. If target TCI state is not in the active TCI state list for PDSCH/PDCCH, TCI state switch delay includes SSB measurement delay, and MAC-CE based beam activation is performed.

Without top K beam sweeping, gNB performs beam indication based on predicted RSRP, top 1 beam prediction accuracy can be 10%~20% lower than top K/1 beam prediction accuracy based on evaluation results in SI phase. Beam indication depends on gNB, then target TCI may be known TCI, target TCI can also be unknown TCI if it is not measured within 1280 ms.

For unknown target TCI indication, only MAC-CE based beam activation is workable, unknown target TCI needs to be measured after MAC-CE activation command to acquire QCL information. For a specific UE, only one unknown target TCI needs to be measured after inference, additional RS measurement overhead for UE is small. RS overhead for gNB is the same as with top K beam sweeping, since in the worst case gNB needs to transmit aperiodic RS in set A after each inference.

Between inference and new target TCI state applies, unknown TCI state switching delay may be included. As defined in 38.133, if the target TCI state is unknown, upon receiving PDSCH carrying MAC-CE activation command in slot n, UE shall be able to receive UE-dedicated PDCCH/PDSCH with target TCI state of the serving cell on which TCI state switch occurs at the first slot that is after slot n+ THARQ + + (TL1-RSRP +TOuk\*(Tfirst-SSB+ TSSB-proc)) / *NR slot length,* where

- T L1-RSRP = 0 in FR1 or when the TCI state switching not involving QCL-TypeD in FR2. Otherwise,

- T L1-RSRP is the time for Rx beam refinement in FR2, defined as

- TL1-RSPR\_Measurement\_Period\_SSB for SSB as specified in clause 9.5.4.1,

- with the assumption of M=1

- with TReport = 0

- TL1-RSRP\_Measurement\_Period\_CSI-RS for CSI-RS as specified in clause 9.5.4.2

- CSI-RS based L1-RSRP measurement only apply for TCI state switch when source RS is associated with serving cell

- configured with higher layer parameter *repetition* set to ON

- with the assumption of M=1 for periodic CSI-RS

- for aperiodic CSI-RS if number of resources in resource set at least equal to *MaxNumberRxBeam*

- with TReport = 0

- TOuk = 1 for CSI-RS based L1-RSRP measurement, and 0 for SSB based L1-RSRP measurement when TCI state switching involves QCL-TypeD

- TOuk = 1 when TCI state switching involves other QCL types only

- Tfirst-SSB is time to first SSB transmission after L1-RSRP measurement when TCI state switching involves QCL-TypeD;

- Tfirst-SSB is time to first SSB transmission after MAC CE command is decoded by the UE for other QCL types;

Without top K beam sweeping, L1-RSRP measurement and SSB measurement are also needed for CSI-RS beam, the total delay is similar to that of with top K beam sweeping procedure.

Based on above analysis, top K beam sweeping procedure introduces some RS measurement overhead but also provides higher beam prediction accuracy, the delay between inference and new target TCI state application is similar to that of without top K beam sweeping procedure. Thus top K beam sweeping procedure can be introduced. Whether top K beam sweeping procedure is enabled is configurable by gNB.

**Proposal 13: Top K beam sweeping procedure can be introduced and is configurable by gNB.**

One problem to configure Top K beam sweeping is Top K beams often change after each inference. If periodic/semi-persistent CSI report is used for Top K beam sweeping, set A is associated with the CSI report. A DL signaling is needed to indicate Top K beam subset for measurement before each report, signaling overhead is large. If aperiodic CSI report is used for Top K beam sweeping, one trigger state can be associated with multiple CSI report and multiple resource sets. Since the maximum number of trigger states for aperiodic CSI reporting and maximum number of resource sets associated with one trigger state is limited, it is difficult to configure all combinations of Top K beams and also occupies too many trigger states and resource sets oriented for other purposes. The method to indicated all combinations of Top K beams for aperiodic CSI reporting with low signaling overhead needs further discussion.

**Proposal 14: The indication of Top K beam set with low signaling overhead needs further discussion.**

### 3.1.2 configuration of set B for BM-Case 2

For NW-sided model for BM-Case 2, if set B equals to set A, as shown in Fig.2, sliding measurement window is not considered. gNB does not transmit set B in prediction instances, which can reduce gNB RS overhead. For periodic/semi-persistent inference report, CSI report framework needs enhancement to configure a resource set transmission pattern like a bundle.

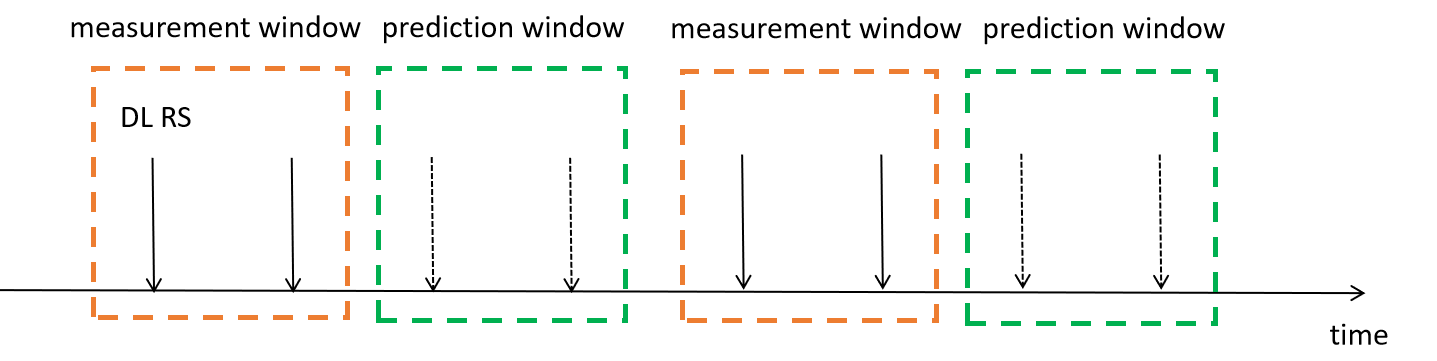


Fig. 2 measurement window for BM-Case 2, set B = set A

**Proposal 15: For NW-sided model, for BM-Case 2 without sliding measurement window, CSI report framework needs enhancement to configure the resource set for measurement.**

If set B is subset of set A, sliding measurement window is considered and determined by gNB. For periodic/semi-persistent set B, gNB can configure to report after each measurement instance, the configuration of measurement window is not needed.

**Proposal 16: For NW-sided model, for BM-Case 2 with sliding measurement window, measurement window is determined by gNB.**

### 3.1.3 reporting content

In RAN1#117 meeting, following agreement regarding reporting content for inference at NW-sided model has been achieved.

Agreement

For NW-sided model, for inference report, at least for BM-Case 1, the content in a beam report in L1 signaling, support

* L1-RSRPs and corresponding beam information of Top M beam(s) with largest M measured value(s) of L1-RSRP(s) of a measurement resource set, where M is configured by gNB
* If M = the size of the measurement resource set, the content is all L1-RSRPs and one beam index (i.e., CRI/SSBRI) for the largest measured value of L1-RSRP of a measurement resource set
* FFS: L1-RSRPs and corresponding beam information of up to M beams within X dB gap to the largest measured value of L1-RSRP, X and M are configured by gNB, and whether/how to report number of reported beams
* FFS on the maximum value of M (where M can be larger than 4) based on UE capability (M may or may not be different for different reporting contents)
* FFS on beam information
* Note: Purpose, such as above “For NW-sided model, for inference report, at least for BM-Case 1”, will not be specified in RAN 1 specifications

If L1-RSRPs of Top M beams are reported and M<the size of the measurement resource set, beam information can be bitmap. Since M is a fixed value configured by gNB, the drawback is some beams with high L1-RSRP beside Top M beams are omitted if M is small. Some beams with low L1-RSRP inside Top M beams are reported if M is large, resulting in waste of UL resource.

To remain beams with high L1-RSRP in report content and save UL resource, one flexible method is to report L1-RSRPs and corresponding beam information of beams within X dB gap to the largest measured L1-RSRP. In this case, the payload size of reporting content may change. Similar to two part report of CSI typeI/typeII codebook, reporting content can be divided into part I and part II, part I includes beam information (e.g. bitmap) with fixed size, part II includes differential L1-RSRPs of variable number of beams. The UL resource allocation for variable payload size needs further discussion.

**Proposal 17: For NW-sided model, for inference report, at least for BM-Case 1, if the content in a beam report in L1 signaling is L1-RSRPs and corresponding beam information of Top M beam(s) with largest M measured value(s) of L1-RSRP(s) and M < the size of the measurement resource set, beam information is bitmap.**

**Proposal 18: For NW-sided model, for inference report, at least for BM-Case 1, the content in a beam report in L1 signaling, support:**

* **L1-RSRPs and corresponding beam information of beams within X dB gap to the largest measured value of L1-RSRP, X is configured by gNB.**
  + **beam information is bitmap**
  + **UL resource allocation for variable number of reported beams needs further discussion**

For NW-sided model for BM-Case 2, if set B is subset of set A, sliding measurement window is considered. For periodic/semi-persistent inference report, UE reports measurement results of one instance/sample in each report with low latency. For aperiodic inference report, UE reports measurement results of multiple instances/sample in one report. The reported beams for each instance are determined as BM-Case 1. In this case, temporal compression for overhead reduction is reasonable. In reporting content, bitmap can be used as beam information to indicate reported beams over multiple instances, the second part is differential L1-RSRPs of variable number of beams.

**Proposal 19: For NW-sided model, for inference report for BM-Case 2, temporal compression for overhead reduction is supported at least for aperiodic report. Beam information can be bitmap.**

### 3.1.4 TCI indication for beam in set A but not in set B

After inference, gNB may choose a beam in set A but not in set B. In the following, we discuss the TCI indication in this case.

One possible way is set A is configured and transmitted for monitoring, or top K beam sweeping is performed. Then TCI indication is associated with measured RS resource and can reuse legacy unify TCI framework.

Another possible way is gNB directly activate an unknown TCI state based on predicted RSRP, TCI indication is associated with virtual RS without transmission. Here, virtual RS in set A is more like aperiodic resource with resource configuration but not transmitted. Based on existing unknown TCI activation timeline, gNB should trigger transmission of this virtual RS once for aperiodic measurement at UE to obtain QCL parameters.

Above all, the RS associated with TCI indication should be measured at least once before TCI application. The UE can not consider a TCI indication associated with virtual RS without transmission as known TCI state since QCL parameters are unknown. This applies for both NW-sided model and UE-sided model.

**Proposal 20: The RS associated with TCI indication should be measured at least once before TCI application. TCI indication associated without RS in set A is not supported.**

## Model inference at UE side

### 3.2.1 necessity of Top-K beam sweeping procedure

The model inference procedure of Tx beam prediction at UE side is shown in Fig. 3.

* Firstly, gNB transmits sparse Tx beams in Set B, where Set B beam pattern may be fixed or variable for different UEs. The detailed configuration of beams in Set B is up to gNB or reported by UE.
* In step 2, UE inputs L1-RSRP of measured beams into AI model and outputs index and L1-RSRP of Top K best beams among all beams.
* In step 3, UE reports the index of predicted Top K beams, the corresponding L1-RSRP can also be reported optionally.
* In step 4, gNB jointly sweeps Top K beams and all Rx beams. UE measures L1-RSRP of Top K best beams.
* In step 5, UE reports the index and L1-RSRP of practical Top 1 best beam or part of Top K best beams. Legacy beam reporting mechanism is reused.
* In step 6, gNB indicates the beam for DL data transmission. Legacy beam indication mechanism is reused.

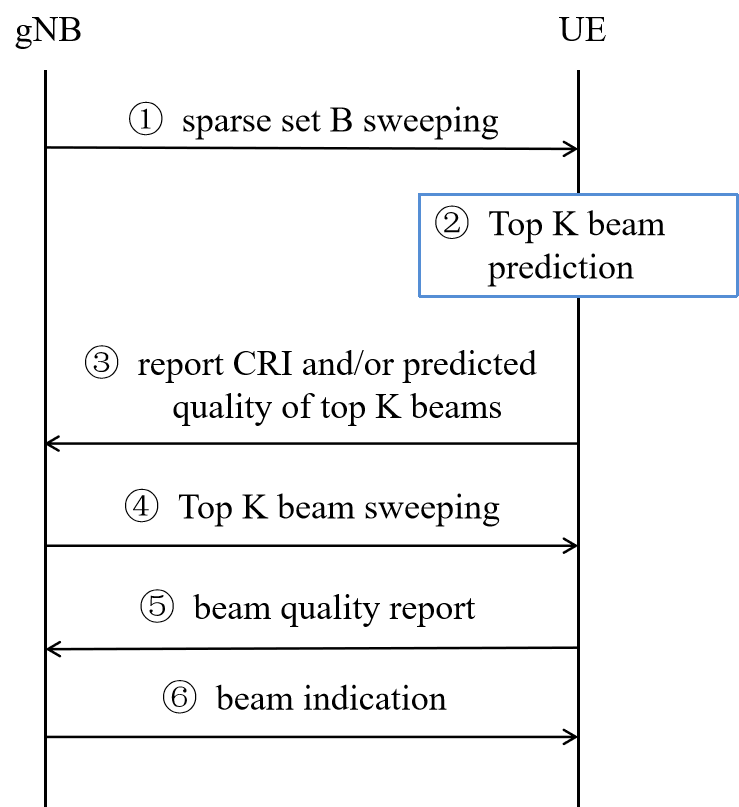


Fig. 3 inference at UE side

Similar as inference at NW side, Top K beam sweeping procedure is configurable by gNB.

### 3.2.2 configuration of set A and set B for BM Case-1

For UE sided model, gNB needs to configure set B for measurement and configure set A for prediction and reporting. In RAN1 116b meeting, five options are provided to configure the association of beams within set B and beams within set A for inference results reporting.

* + Option 1: one CSI-ResourceConfigId is configured for Set B
    - FFS: how UE can determine the information about set A
  + Option 2: one *CSI-ResourceConfigId* is configured for both Set A and Set B
  + Option 3: two *CSI-ResourceConfigId* s are configured for Set A and Set B separately
  + Option 4: one *CSI-ResourceConfigId* is configured for Set B, Set A is configured using separate resource set(s) other than that represented by *CSI-ResourceConfigId*
  + Option 5: separate *CSI-ReportConfig* for Set A and Set B

We prefer union design of inference results reporting and data collection for training and monitoring, instead of one signaling structure for training and monitoring and another signaling structure for inference. The major difference is set A needs measurement for training but does not need measurement for inference.

Since configuring set A is necessary for data collection for training and monitoring, we prefer to configure set A firstly and indicate set B with association of two set, thus option 1 is deprioritized. With option 1, it is also hard to indicate the association of set B and set A without any information/configuration of set A, and hard to perform TCI indication associated without RS in set A.

Option 2 and 3 are reasonable, separate resource sets for Set A and Set B are necessary to distinguish resource for measurement and resource for prediction. Differently from training in section 2.2.2, if gNB does not transmit Set A actually, e.g. when monitoring has not been triggered, Set A can be a virtual set. RS in Set A has RS ID and resource configuration but is not transmitted by gNB. When aperiodic transmission of some RS in set A is triggered for TCI indication as mentioned in section 3.1.4, gNB transmits these RS once. RS ID or bitmap is considered for indication of association of set A and set B.

Option 4 is similar to option 3 but the benefit is not clear.

With option 5, one *CSI-ReportConfig* only includes set B for measurement, one *CSI-ReportConfig* only includes Set A for prediction. UE needs to link two *CSI-ReportConfig* together to understand which is the set of beams for reporting. Comparing with Option 2 and 3, option 5 seems more complicated and without additional benefit.

**Proposal 21: For UE-sided model at least for BM Case-1, *CSI-ReportConfig* is used for the configuration of inference results reporting**

* **FFS on the details in the *CSI-ReportConfig*, at least considering:**
  + **Option 2: one *CSI-ResourceConfigId* is configured for both Set A and Set B**
    - **separate resource sets are configured for Set A and Set B in *CSI-ResourceConfig***
    - **Set A may be a virtual set including RS ID and resource configuration, RS in Set A is not transmitted unless triggered by gNB**
    - **indication of association of Set A and Set B can be based on RS ID or bitmap**
  + **Option 3: two *CSI-ResourceConfigId* s are configured for Set A and Set B separately**
    - **Set A may be a virtual set including RS ID and resource configuration, RS in Set A is not transmitted unless triggered by gNB**
    - **indication of association of Set A and Set B can be based on RS ID**

### 3.2.3 configuration of set B for BM-Case 2

For UE-sided model for BM-Case 2, if set B equals to set A, sliding measurement window is not considered. For periodic/semi-persistent inference report, CSI report framework needs enhancement to configure a resource set transmission pattern like a bundle.

**Proposal 22: For UE-sided model, for BM-Case 2 without sliding measurement window, CSI report framework needs enhancement to configure the resource set for measurement.**

If set B is subset of set A, sliding measurement window is considered and determined by UE. For periodic/semi-persistent inference report, gNB only configures prediction instances, the configuration of measurement window is not needed.

**Proposal 23: For UE-sided model, for BM-Case 2 with sliding measurement window, measurement window is determined by UE.**

### 3.2.4 inference results reporting for BM-Case 2

In RAN1#117 meeting, FL provided following proposal about configuration of inference results reporting for BM-Case 2.

Proposal:

For UE-side AI/ML model inference, for BM-Case2, to report inference results of N future time instance(s) in one report

* Each of the N future time instance(s) consists of P (P≥1) consecutive slots
  + FFS: How to determine P
* For the reference time to determine the earliest time instance from the N future time instance(s), consider the following options:
  + Option 1: Based on the time domain resource for the report
  + Option 2: Based on the CSI reference resource corresponding to the report
  + Option 3: Based on the transmission occasion of the CSI-RS/SSB resource in Set B for the report
  + FFS: whether the above options are also applicable to the time instance(s) other than the earliest one
  + FFS: If N>1, whether the time domain separation between two adjacent time instance(s) from the N future time instances are the same
* FFS: How to define measurement window(s) for the inference results of the N future time instance(s)

Firstly, the meaning of P needs clarification. We think P is the time interval of N future time instances instead of duration of future time instances. P is configured by gNB based on the meta info of AI model.

Among three options of the reference time to determine the N future time instances, referring to the configuration method of prediction instances for CSI prediction in R18 MIMO, option 1 is prioritized. Option 1 is clear and intuitive to determine the UL slot for the report compared with option 2 and option 3. In detail, the earliest predicted instance can be determined by UL slot for the report and a time offset. The rest predicted instances can use the earliest predicted instance as reference time and be determined by N and P. Option 1 is suitable for the case measured time instances and predicted time instances have the same or different time interval.

With option 2, CSI reference resource is determined indirectly based on UL slot of the report. With option 3, for UE-side model, the transmission occasion of RS is known by configuration, but the measured occasion is up to UE. Thus the reference time can not be based on the measured occasion. If the most recent occasion before CSI reference resource is used as the reference time to determine predicted time instances, it is still an indirect way.

**Proposal 24: For UE-side AI/ML model inference, for BM-Case2, to report inference results of N future time instance(s) in one report**

* **The time interval of N future time instance(s) is P (P≥1) slots, P is configured by gNB**
* **For the reference time to determine the earliest time instance from the N future time instance(s), option 1 is supported:**
  + **Option 1: Based on the time domain resource for the report + offset**
    - **the time instance(s) other than the earliest one is determined by slot of report + offset + N + P**
    - **measured time instances and predicted time instances may have the same or different time interval**
  + **~~Option 2: Based on the CSI reference resource corresponding to the report~~**
  + **~~Option 3: Based on the transmission occasion of the CSI-RS/SSB resource in Set B for the report~~**

### 3.2.5 reporting content

For Top K beams determined by predicted L1-RSRP, some beams within them may have been measured in step 1, whether to report measured L1-RSRP instead of predicted L1-RSRP for these beams needs discussion. RAN1#116b has agreed to further study the following options for BM-Case1:

* Option A: Predicted RSRP
* Option B: Predicted RSRP, if the beam is not configured for corresponding measurement, and measured L1-RSRP if the beam is configured for corresponding measurement
* Note: Support both Option A and Option B is not precluded.

With option A, the predicted L1-RSRP can provide the beam quality information to gNB. If all the predicted L1-RSRP of the Top-K beams is low, the gNB may not to continue the following measurements of the Top-K beams, and the gNB can select other Set B beam patterns for more accurate beam prediction.

With option B, the benefit is measurement overhead reduction of step 4, especially when one measured L1-RSRP is larger than all other L1-RSRPs, gNB may directly perform beam indication. Whether the predicted L1-RSRP should be reported can be configured by gNB.

If gNB does not indicate to apply option A or option B, reported RSRP is determined by UE. Then based on the note, UE can choose to report predicted RSRP or measured L1-RSRP for each Top K beams within set B, method to distinguish predicted RSRP and measured L1-RSRP in reporting content needs further discussion.

**Proposal 25: For report content of inference results for UE-sided model for BM-Case 1 whether the predicted L1-RSRP is reported can be configured by gNB, whether/how to differentiate measured L1-RSRP and predicted L1-RSRP needs further discussion.**

### 3.2.6 consistency of NW-side additional condition across training and inference

Regarding the consistency of NW-side additional condition across training and inference for UE-sided model, RAN1 116b meeting agreed to further study following two options:

* Opt1: Based on associated ID (Referring to AI 9.1.3.3)
  + FFS on what can be assumed by UE with the same associated ID across training and inference
  + FFS on how associated ID is introduced, e.g., within CSI framework, or outside of CSI framework
* Opt 2: Performance monitoring based
  + FFS details

Before we agree on option 1, we need to discuss the feasibility and application range of associated ID.

With option 1, NW transmits the data collection related configurations and their associated ID, UE collects the data corresponding to the associated ID, and AI/ML models are trained at UE side based on the collected data corresponding to the associated ID. Thus, associated ID of a model or function is related to NW side additional conditions, these additional conditions may include NW side proprietary information, e.g. 3dB beamwidth, beam boresight directions.

After model training and UE accesses to another gNB with different NW-side additional conditions, UE needs to report associated ID, and another gNB should identify the NW side additional conditions behind associated ID. So that another gNB can judge whether it can provide the same additional conditions as the associated ID.

We can conclude the application of option 1 is based on the premise that different gNB can identify the additional conditions behind associated ID. However, this may be difficult among different gNB vendors. Thus, different gNB vendors should use different ID set. If one gNB receives an associated ID from a UE and the associated ID does not belong to this gNB vendor’s associated ID set, gNB does not transmit inference configuration for this AI model.

For the same gNB vendor, within an ID set, an ideal status is using global ID for associated ID, and the base stations of the same gNB vendor align the additional conditions behind associated ID offline. It is also possible that associated ID is configured per area like PCI. One associated ID in different area may correspond to different NW side additional conditions, the base stations of the same gNB vendor can interact the additional conditions behind associated ID.

**Proposal 26: Regarding to the application range of associated ID,**

* **different gNB vendors use different associated ID set.**
* **for the same gNB vendor, associated ID is global ID or is configured per area.**

Regarding to the UE assumption behind the same associated ID, based on our proposal 21, different gNB vendors do not need to align additional conditions behind associated ID since they use different associated ID set. Among base stations of the same gNB vendor, UE assumptions including the same transmission filter, the same order of resources (corresponding to beams) for Set A/B, the same Tx power of gNB, the same antenna height and down tilt and deployment scenarios, can be assumed with the same associated ID. These NW-side additional conditions behind associated ID can be aligned without exposing proprietary information.

**Proposal 27: Regarding to the UE assumptions behind the same associated ID,**

* **different gNB vendors do not use the same associated ID.**
* **for the same gNB vendor, UE assumptions including the same transmission filter, the same order of resources (corresponding to beams) for Set A/B, the same Tx power of gNB, the same antenna height and down tilt, the same deployment scenarios.**

Before monitoring, gNB can preliminary guarantee consistency of NW-side additional condition without exposing proprietary information based on associated ID to reduce measurement and reporting overhead of monitoring. For example, if gNB can not provide the same transmission filter of set A beams as associated ID of a model, this model will not be activated.

Since the configuration of set A and set B reuses CSI reporting framework, associated ID can also be introduced within CSI framework, e.g. per CSI-ReportConfig. Since one model collects measurement results of set A and B from the same gNB or multiple gNB, the additional conditions behind set A and B are the same under each gNB, the associated ID for set A and B are the same under each gNB. There is no need to configure associated ID with granularity smaller than resource set.

If the gNB can provide the same additional conditions behind associated ID of a model, there are still other NW-side additional conditions not guaranteed by associated ID, then performance monitoring is still a baseline to guarantee the performance during inference.

**Proposal 28: For the consistency of NW-side additional condition across training and inference for UE-sided model for BM-Case 1 and BM Case 2, option 1 and option 2 are supported:**

* **Opt1: Based on associated ID (Referring to AI 9.1.3.3)**
  + **associated ID is introduced within CSI framework and configured per CSI-ReportConfig.**
* **Opt 2: Performance monitoring based**

# Discussion on specification impact of monitoring

Model monitoring can timely identify model performance degradation and performs model updating/switching/fallback to guarantee good system performance, which is important for model life cycle management. In this section, we discuss the monitoring mechanism, KPI and benchmark.

## monitoring mechanism

For NW-side model, NW can calculate the performance metric, make decision and perform model selection/activation/ deactivation/switching/fallback directly, it is natural that NW-side monitoring is applied to NW-side model. There is no need to define KPI to give more flexibility to NW-side decision. For example, NW may make decision considering prediction accuracy, RSRP difference and throughput jointly.

**Proposal 29: NW-side monitoring of NW-side AI/ML model is supported. The KPI is up to gNB.**

For UE-side model, the following model monitoring mechanisms have been captured in TR 38.843, and Type 1 performance monitoring has been agreed in RAN1#117 meeting.

- Type 1 performance monitoring:

- Configuration/Signalling from gNB to UE for measurement and/or reporting

- UE may have different operations

- Option 1 (NW-side performance monitoring): UE sends reporting to NW (e.g., for the calculation of performance metric at NW)

- Option 2 (UE-assisted performance monitoring): UE calculates performance metric(s), either reports it to NW or reports an event to NW based on the performance metric(s)

- Indication from NW for UE to do LCM operations

- Note: At least the performance and reporting overhead of model monitoring mechanism should be considered

- Type 2 performance monitoring:

- Indication/request/report from UE to gNB for performance monitoring

- Note: The indication/request/report may be not needed in some case(s)

- Configuration/Signalling from gNB to UE for performance monitoring measurement and/or reporting

- If it is for UE side model monitoring, UE makes decision(s) of model selection/activation/ deactivation/switching/fallback operation

RAN2#125bis has approved that for function based LCM of UE-sided model, the “network decision, network-initiated” AI/ML management is supported as a baseline. “UE autonomous, decision reported to the network” can be considered further, while “UE-autonomous, UE’s decision is not reported to the network” is not considered for Rel-19.

We can conclude that UE autonomous AI/ML management corresponding to Type 2 performance monitoring has lower priority, and it is necessary to report UE decision to the network to let network know the change of AI function and configuration.

For Type 2 monitoring, NW may configure a threshold criterion to facilitate UE to make decision. For example, if beam prediction accuracy of active model is larger than threshold, the active model is available and switching/fallback operation is unnecessary. If beam prediction accuracy of active model is smaller than threshold, only inactive model with beam prediction accuracy larger than threshold is considered as candidate model for model switching.

Similar as the agreement of RAN2, the decision of model selection/activation/deactivation/switching/fallback operation and the configuration of target new model should be reported to the network, and the decision can apply after UE receiving the acknowledgement from the network. For example, if model switching results in configuration update of set B, UE reports decision and new set B configuration to NW. If all other UEs use origin set B and NW does not want to configure new set B, NW can reject model switching. If UE does not report decision to NW and directly performs model switching, it may cause the misunderstanding of set B configuration between NW and UE.

**Proposal 30: Regarding Type 2 monitoring of UE-side AI/ML model, NW may configure a threshold criterion to facilitate UE to perform model monitoring. The specification impact of decision and configuration reporting, and decision acknowledgement mechanism are considered.**

## reporting signaling

Data collection for monitoring is time-critical and will sustain for a long period until AI based beam management fallback to legacy beam management, thus existing periodic/semi-persistent L1 CSI reporting framework can be reused.

**Proposal 31: Regarding signaling of data collection for monitoring, L1 signaling is supported.**

## monitoring KPI and benchmark

For Type 1 option 1 monitoring, UE sends a report to NW for the calculation of performance metric at NW. There is no need to define KPI to give more flexibility to NW-side decision. NW can make decision based on global information. For example, although beam prediction accuracy reduces 10%, but throughput reduction is slight or RSRP difference between genie-aided Top 1 beam and predicted Top 1 beam is small, NW can decide to remain current AI model instead of model switching based on a fixed beam prediction accuracy threshold.

The UE report is a single sample-based report. The report content can be measurement results of resource set for monitoring, e.g., L1-RSRP and/or Top 1 RS ID. In this case, the benefit is gNB can calculate desired KPI flexibly, such as Top-1 or Top-3/1 or Top-K/1 prediction accuracy, set B or Top K RSRP difference, but L1 reporting overhead is larger. Another possible way is gNB configures to report whether Top-K/1 beam prediction is correct per sample, then K is a fixed value, L1 reporting payload is small since report content has one bit.

**Proposal 32: Regarding Type 1 option 1 monitoring, of UE-side AI/ML model, KPI is up to gNB. UE report is a single sample-based report. The report content can be measurement results of resource set for monitoring, e.g., L1-RSRP and/or Top 1 RS ID, or Top-K/1 prediction accuracy per sample.**

For Type 1 option 2 monitoring, we think beam prediction accuracy can be used as model monitoring KPI, since it directly reflects whether AI model can find Top 1 beam and it is directly proportional to throughput. If monitoring KPI is defined as whether Top-K/1 beam prediction is accurate per sample, KPI is reported frequently. If monitoring KPI is defined as statistical results on Top-K/1 beam prediction accuracy over N instances, NW may configure a threshold criterion or an event to facilitate UE to judge model performance. KPI is reported when it is less than threshold or when the event is satisfied, then reporting has less frequency. As a new report quantity, the quantification method of Top-K/1 beam prediction accuracy needs to be discussed. Referring to BFD triggering event, the event can be defined as consecutive N1 times of KPI/BLER less than threshold, or accumulated N1 times of KPI/BLER less than threshold within a time period.

If model switching may result in inference configuration update, the KPI and/or the configuration of inactive models should be reported to the network to facilitate NW side decision.

**Proposal 33: Regarding Type 1 option 2 monitoring of UE-side AI/ML model, KPI can be single sample-based Top-1/K beam prediction accuracy or multiple sample-based Top-1/K beam prediction accuracy.**

**Proposal 34: Regarding Type 1 option 2 monitoring of UE-side AI/ML model, NW may configure a threshold criterion or event to facilitate UE to perform model monitoring. The event can be defined as consecutive N1 times of KPI/BLER less than threshold, or accumulated N1 times of KPI/BLER less than threshold within a time period.**

Based on the agreement in RAN1#112b, the following alternatives are considered as the benchmark/reference (if applicable) for monitoring performance comparison:

· Alt.1: The best beam(s) obtained by measuring beams of a set indicated by gNB (e.g., Beams from Set A)

o FFS: gNB configures one or multiple sets for one or multiple benchmarks/references

· Alt.4: Measurements of the predicted best beam(s) corresponding to model output (e.g., Comparison between actual L1-RSRP and predicted RSRP of predicted Top-1/K Beams)

· FFS:

o Alt.3: The beam corresponding to some or all the indicated/activated TCI state(s)

· Other alternative is not precluded.

Alt1 implies that different benchmarks can be derived and used together, such as upper bound of performance (Top 1 genie-aided beam from Set A), lower bound of performance for fallback (best beam from Set B) and monitoring KPI of multiple inactive AI models. With Alt1, UE or NW can simultaneously monitor the performance difference among active model and inactive model and non-AI beam management, which facilitates UE or NW to make monitoring decision such as finding target inactive model with better performance than active model.

For Type 2 and Type 1 option 2 monitoring of UE-side AI/ML model, monitoring mechanism of multiple benchmarks needs design. For example, similar as BFR mechanism, if monitoring KPI of active model is lower than threshold by N times within timer, UE decides to deactivate active model or report this event to NW. Multiple timers for monitoring active model and inactive models can be parallel or sequential depending on UE capability.

For Alt4, evaluation is needed to confirm whether RSRP difference can reflect beam prediction accuracy and throughput variation.

**Proposal 35: The best beam(s) obtained by measuring beams of a set indicated by gNB is considered as the benchmark/reference for monitoring performance comparison:**

**o gNB configures one or multiple sets for one or multiple benchmarks/references**

**Proposal 36: For Type 2** and **Type 1 option 2 monitoring of UE-side AI/ML model, monitoring mechanism of multiple benchmarks needs discussion to facilitate UE to perform model monitoring.**

# Conclusion

In this contribution, we discussed the potential specification impact of AI based beam management during different LCM operation, and the following proposals are made.

**Proposal 1: L1 signaling is supported for NW-sided training data collection.**

**Proposal 2: For NW-sided model, for data collection for training, at least for BM-Case1, option 1 and 2 are supported for the report contents:**

* **Opt 1: L1-RSRPs from RS resources from one set of beams configured to UE**
  + **for higher layer report, beam information is CRI of RS with largest L1-RSRP**
  + **for L1 signaling, beam information can be bitmap**
* **Opt 2: L1-RSRPs from RS resources from one set of beams configured to UE, and beam information of Top K from RS resources from another set of beams**
  + **K is configured by RRC, value of K is 1~10**
  + **for higher layer report, beam information is CRI of RS with largest L1-RSRP in measurement set and CRI of Top K RS in prediction set**
  + **for L1 signaling, beam information is bitmap and CRI of Top K RS in prediction set**

**Proposal 3: If L1 signaling is applied for NW-sided training data collection, further discuss reporting all or a subset of L1-RSRPs from the resource set, at least including data omission, e.g., L1-RSRP(s) higher than a threshold.**

**Proposal 4: Regarding data collection for NW-side model, following options can be considered for the configuration of Set A:**

* **Option 1: The UE capability of the maximum number of the RS per resource set for the RSRP measurement can be enhanced.**
* **Option 2: Multiple resource sets can be supported in one *CSI-ResourceConfig***

**Proposal 5: Regarding training data collection for NW-side model, the configuration of multiple set B or multiple Set B patterns can be supported.**

**Proposal 6: Rx beam assumption for a measurement report can be up to gNB implementation. If quasi-optimal Rx beam assumption applies, an indication may be needed in CSI-ReportConfig to align the Rx beam assumption between gNB and UE.**

**Proposal 7: Regarding data collection for UE-side model for training, UE can request for preferred set B.**

**Proposal 8: Regarding data collection for UE-side model for training, set B as AI model input is determined by gNB.**

**Proposal 9: Regarding data collection for UE-side model, the configuration method of set A for NW-side model can be reused.**

* **Option 1: The UE capability of the maximum number of the RS per resource set for the RSRP measurement can be enhanced.**
* **Option 2: Multiple resource sets can be supported in one *CSI-ResourceConfig***

**Proposal 10: For UE-sided model at least for BM Case-1, following two options are supported for the configuration of training data collection.**

* + **Option 1: one *CSI-ResourceConfigId* is configured for both Set A and Set B**
    - **separate resource sets are configured for Set A and Set B in *CSI-ResourceConfig***
  + **Option 2: two *CSI-ResourceConfigId* s are configured for Set A and Set B separately**

**Proposal 11: Regarding to option 1 of training data collection for UE-sided model, indication of association of Set A and Set B can be based on RS ID or bitmap.**

**Proposal 12: Regarding to option 2 of training data collection for UE-sided model, indication of association of Set A and Set B can be based on RS ID.**

**Proposal 13: Top K beam sweeping procedure can be introduced and is configurable by gNB.**

**Proposal 14: The indication of Top K beam set with low signaling overhead needs further discussion.**

**Proposal 15: For NW-sided model, for BM-Case 2 without sliding measurement window, CSI report framework needs enhancement to configure the resource set for measurement.**

**Proposal 16: For NW-sided model, for BM-Case 2 with sliding measurement window, measurement window is determined by gNB.**

**Proposal 17: For NW-sided model, for inference report, at least for BM-Case 1, if the content in a beam report in L1 signaling is L1-RSRPs and corresponding beam information of Top M beam(s) with largest M measured value(s) of L1-RSRP(s) and M < the size of the measurement resource set, beam information is bitmap.**

**Proposal 18: For NW-sided model, for inference report, at least for BM-Case 1, the content in a beam report in L1 signaling, support:**

* **L1-RSRPs and corresponding beam information of beams within X dB gap to the largest measured value of L1-RSRP, X is configured by gNB.**
  + **beam information is bitmap**
  + **UL resource allocation for variable number of reported beams needs further discussion**

**Proposal 19: For NW-sided model, for inference report for BM-Case 2, temporal compression for overhead reduction is supported at least for aperiodic report. Beam information can be bitmap.**

**Proposal 20: The RS associated with TCI indication should be measured at least once before TCI application. TCI indication associated without RS in set A is not supported.**

**Proposal 21: For UE-sided model at least for BM Case-1, *CSI-ReportConfig* is used for the configuration of inference results reporting**

* **FFS on the details in the *CSI-ReportConfig*, at least considering:**
  + **Option 2: one *CSI-ResourceConfigId* is configured for both Set A and Set B**
    - **separate resource sets are configured for Set A and Set B in *CSI-ResourceConfig***
    - **Set A may be a virtual set including RS ID and resource configuration, RS in Set A is not transmitted unless triggered by gNB**
    - **indication of association of Set A and Set B can be based on RS ID or bitmap**
  + **Option 3: two *CSI-ResourceConfigId* s are configured for Set A and Set B separately**
    - **Set A may be a virtual set including RS ID and resource configuration, RS in Set A is not transmitted unless triggered by gNB**
    - **indication of association of Set A and Set B can be based on RS ID**

**Proposal 22: For UE-sided model, for BM-Case 2 without sliding measurement window, CSI report framework needs enhancement to configure the resource set for measurement.**

**Proposal 23: For UE-sided model, for BM-Case 2 with sliding measurement window, measurement window is determined by UE.**

**Proposal 24: For UE-side AI/ML model inference, for BM-Case2, to report inference results of N future time instance(s) in one report**

* **The time interval of N future time instance(s) is P (P≥1) slots, P is configured by gNB**
* **For the reference time to determine the earliest time instance from the N future time instance(s), option 1 is supported:**
  + **Option 1: Based on the time domain resource for the report + offset**
    - **the time instance(s) other than the earliest one is determined by slot of report + offset + N + P**
    - **measured time instances and predicted time instances may have the same or different time interval**
  + **~~Option 2: Based on the CSI reference resource corresponding to the report~~**
  + **~~Option 3: Based on the transmission occasion of the CSI-RS/SSB resource in Set B for the report~~**

**Proposal 25: For report content of inference results for UE-sided model for BM-Case 1 whether the predicted L1-RSRP is reported can be configured by gNB, whether/how to differentiate measured L1-RSRP and predicted L1-RSRP needs further discussion.**

**Proposal 26: Regarding to the application range of associated ID,**

* **different gNB vendors use different associated ID set.**
* **for the same gNB vendor, associated ID is global ID or is configured per area.**

**Proposal 27: Regarding to the UE assumptions behind the same associated ID,**

* **different gNB vendors do not use the same associated ID.**
* **for the same gNB vendor, UE assumptions including the same transmission filter, the same order of resources (corresponding to beams) for Set A/B, the same Tx power of gNB, the same antenna height and down tilt, the same deployment scenarios.**

**Proposal 28: For the consistency of NW-side additional condition across training and inference for UE-sided model for BM-Case 1 and BM Case 2, option 1 and option 2 are supported:**

* **Opt1: Based on associated ID (Referring to AI 9.1.3.3)**
  + **associated ID is introduced within CSI framework and configured per CSI-ReportConfig.**
* **Opt 2: Performance monitoring based**

**Proposal 29: NW-side monitoring of NW-side AI/ML model is supported. The KPI is up to gNB.**

**Proposal 30: Regarding Type 2 monitoring of UE-side AI/ML model, NW may configure a threshold criterion to facilitate UE to perform model monitoring. The specification impact of decision and configuration reporting, and decision acknowledgement mechanism are considered.**

**Proposal 31: Regarding signaling of data collection for monitoring, L1 signaling is supported.**

**Proposal 32: Regarding Type 1 option 1 monitoring, of UE-side AI/ML model, KPI is up to gNB. UE report is a single sample-based report. The report content can be measurement results of resource set for monitoring, e.g., L1-RSRP and/or Top 1 RS ID, or Top-K/1 prediction accuracy per sample.**

**Proposal 33: Regarding Type 1 option 2 monitoring of UE-side AI/ML model, KPI can be single sample-based Top-1/K beam prediction accuracy or multiple sample-based Top-1/K beam prediction accuracy.**

**Proposal 34: Regarding Type 1 option 2 monitoring of UE-side AI/ML model, NW may configure a threshold criterion or event to facilitate UE to perform model monitoring. The event can be defined as consecutive N1 times of KPI/BLER less than threshold, or accumulated N1 times of KPI/BLER less than threshold within a time period.**

**Proposal 35: The best beam(s) obtained by measuring beams of a set indicated by gNB is considered as the benchmark/reference for monitoring performance comparison:**

**o gNB configures one or multiple sets for one or multiple benchmarks/references**

**Proposal 36: For Type 2** and **Type 1 option 2 monitoring of UE-side AI/ML model, monitoring mechanism of multiple benchmarks needs discussion to facilitate UE to perform model monitoring.**

# References

1. Chair notes RAN1#117 eom1, RAN1#117, Fukuoka City, Fukuoka, Japan, May 20th – 24th, 2024.