3GPP TSG RAN WG1 #112bis-e R1-23xxxxx

e-Meeting, April 17th – April 26th, 2023

Source: Moderator (OPPO)

Title: Summary#4 for other aspects on AI/ML for beam management

Agenda Item: 9.2.3.2

Document for: Discussion and Decision

# Introduction

The Rel-18 WID of AI/ML for NR Air Interface focuses on a subset of three typical use cases:

1. CSI feedback enhancement
2. Beam management
3. Positioning accuracy improvement.

This document focuses on the other aspects of AI/ML for beam managements, including representative sub use cases and potential specification impact.

Regarding the file names, companies are encouraged to follow the guidance of R1-2203012 (Page 16) as below:

|  |
| --- |
| * + - To avoid ending-up with too long file names and downloading/opening issues, the following naming convention is recommended:       * Keep the previous company’s name (only the most recent one) in the filename, e.g.         + 5/Summary-1-v000-Moderator (HW)         + 5/Summary-1-v001-LG         + 5/Summary-1-v002-LG-CATT         + 5/Summary-1-v003-CATT-vivo         + 5/Summary-1-v004-Moderator(HW)       * It helps identifying on which previous version your input is based on and solve any crossing emails issue. Note the use of 3digit version numbers in the file names. |

In the following sections, the company proposals are summarized, and offline proposals are drafted based on company contributions for discussion/input.

# Spec impact of Data collection

## General/common aspects

In previous RAN1 meeting(s), the agreement(s)/conclusion(s) were made as below:

|  |
| --- |
| **RAN1#110**  Agreement  For the data collection for AI/ML model training (if supported), study the following aspects as a starting point for potential necessary specification impact:   * Signaling/configuration/measurement/report for data collection, e.g., signaling aspects related to assistance information (if supported), Reference signals * Content/type of the collected data * Other aspect(s) is not precluded |

The related proposals/ observations related are copied as below:

|  |  |
| --- | --- |
| Huawei[2] | *Proposal 16: For the data collection for model training, study how to enable the UE to measure the Set A with large number of Tx beams which may be restricted by the legacy UE capability on the maximum number of configurable RS resources.* |
| Intel[10] | *Observation 1: The impact of 3GPP specification related procedures for data collection for training as well as inference depends on where the model resides and if training and inferencing is being performed at the same node.*  *Observation 2: Training dataset construction using 3GPP specified measurement and reporting framework may be advantageous for harmonizing deployment of proprietary AI/ML models.* |
| NVIDIA[24] | *Proposal 7: For AI/ML model training for beam prediction in spatial/time domain, study potential specification impact related to training data type/size, training data source determination, and assistance signalling and procedure for training data collection.* |
|  |  |

**Mod’s assessment:** In the last meeting, as triggered by tdocs of NVIDIA/MTK, a proposal was suggested to discuss the requirements of data set (e.g., data size). However, most companies didn’t support it. Thus, moderator has no plan to repeat the same discussion as we can expect the output. The other proposals are quite general or can be discussed in other session (e.g., UE capability). In summary, no discussion point is suggested here.

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |
|  |  |
|  |  |

## Network-side AI model training at NW side

|  |  |
| --- | --- |
| FUTUREWei[1] | *Proposal 6: Regarding the data collection mechanism for NW-side AI/ML model training at NW side, study the following approach as the baseline mechanism (Opt.3 in Proposal 3.2.1 from RAN1#112).*   * *UE reports M4 L1-RSRPs optionally with beam (pair) indicators based on the measurement corresponding to a beam set (e.g., Set B), reports M5 beam (pair) indicators based on the measurement corresponding to another beam set (e.g., Set A), where M4 can be larger than 4*   + *FFS: the range of M4, M5*   + *Other option(s) is not precluded* |
| Huawei[2] | *Proposal 14: Spec impact for training data collection for NW-side AI/ML model should be studied to facilitate initial training and/or on-demand model updating.*  *Proposal 15: Regarding the training data collection mechanism for NW-side AI/ML model trained at NW side, study the following options as a starting point for the contents of collected data*   * *Opt.1: UE sends M1 L1-RSRPs (corresponding to M1 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to a beam set (e.g., Set A), where M1 can be larger than 4*   + *FFS: the range of M1* * *Opt.2: UE sends M2 L1-RSRPs (corresponding to M2 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to a beam set (e.g., Set B), sends M3 L1-RSRPs (corresponding to M3 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to another beam set (e.g., Set A), where M2 and M3 can be larger than 4*   + *FFS: the range of M2, M3* * *Opt.3: UE sends M4 L1-RSRPs (corresponding to M4 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to a beam set (e.g., Set B), sends M5 beams (beam pair) based on the measurement corresponding to another beam set (e.g., Set A), where M4 can be larger than 4*   + *FFS: the range of M4, M5*   *Observation 11: Considering the low frequency of data collection and the tremendous number of sites and UEs to report the data samples, the average air-interface overhead for per UE is negligible.*  *Proposal 17: RAN1 to further study the potential spec impact of data collection from the following aspects:*   * *For reference signal, enhanced RS design can be considered, e.g., RS design for AI/ML specific RSRP measurement and enhancement of RS for improving data sample accuracy* * *For UE measurement/report, new RSRP and/or CRI/SSBRI report behavior can be considered*   + *Data quality indicator can be considered to improve the quality of the collected data samples.* * *For the signaling/configuration, signaling to trigger/configure/request data collection window can be considered*   *Observation 13: For the container of the reported data samples in data collection, L1 signaling has lower latency and is applicable to training and monitoring, while RRC signaling has larger latency and is applicable only to training.*  *Observation 14: For the overhead of the reported data samples in data collection, L1 signaling can be comparable with L3 signaling as the monitoring window with L1 signaling can be triggered with an on-demand manner rather than always-on.*  *Proposal 18: For the potential spec impact of data collection,*   * *Both L1 signaling and RRC signaling can be considered to carry the reported data samples for model training.* * *At least L1 signaling should be considered for model monitoring to enable fast identification of AI/ML model failure.*   *Proposal 19: For the data collection for UE-side/NW-side AI/ML model, study how to indicate the purpose of the RS configurations to differentiate the UE report manners, e.g.,*   * *Differentiate the UE report manners among training, monitoring, and inference.* * *Differentiate the UE report manners between Set A and Set B.* * *Differentiate the UE report manners between AI/ML-based output and legacy measurement report.*   *Proposal 20: For the training/monitoring data collection of AI/ML model at NW side, study the methods to enable UE to feedback the RSRP labels for a subset of all measured beams in Set A to save UE report overhead.*  *Proposal 21: For the training/monitoring data collection of NW-side AI/ML model, the motivation of introducing finer resolution for UE reported RSRP measurement results for labels in Set A may be discussed after being justified in 9.2.3.1.* |
| H3C[3] | *Proposal 3: For the data collection with network-side model, leave the study of RRC signaling based mechanism to trigger/stop/report the AI/ML training data collection to RAN2.*  *Proposal 4: For the data collection with network-side model, study necessity and potential specification impact on the training data format (L1-RSRP, timestamps, RS-indicator, etc.)* |
| ZTE[4] | *Proposal 13: For BM-Case1 and BM-Case2, the collected data for model training includes the measured L1-RSRP, beam ID, corresponding time stamp information, and other assistance information assessed to be necessary in agenda 9.2.3.1.*  *Observation 6: The RS resource transmission for channel measurement can be either be initiated by gNB or requested by UE, and are strongly dependent on the construction of set A/B.*  *Proposal 14: Support to study resource configuration aspects of data collection and associated specification impact with potentially enhanced signaling mechanisms and auxiliary information transmission.*  *Proposal 15: For data collection from UE to NW side, support to study explicit or implicit Rx beam ID reporting method, especially for beam pair prediction.*  *Proposal 16: For data collection from UE to NW side, depending on the model training strategy, model output data can be genie-aided best beam ID from set A, all measurement results of set A or other post-processing of measurement results of set A.*  *Proposal 17: If all measurement results of set A/B need to be reported to gNB, suggest to further study reporting overhead reduction methods, e.g., beam ID can be obtained implicitly from the reporting order of all measured RSRPs.*  *Observation 7: Due to the relatively long data collection duration, the wireless link between gNB and UE may change rapidly, resulting in unguaranteed data sample quality.*  *Proposal 18: For data collection from UE to NW side, study UE-side data filtering and associated reporting mechanism to reduce the reporting overhead.* |
| Vivo[5] | *Proposal 16: Regarding the data collection for AI/ML model training at NW side, study potential specification impact on resource configuration:*  *• Specific beam pair resource configuration for Set A*  *• Enhanced P3+P2 resource configuration that Rx beam assumption of P2 resource measurement is the best Rx beam searched from P3 procedure for performance improvement*  *Proposal 17: Regarding the data collection for AI/ML model training at NW side, study potential specification impact on assistance information:*  *• Proprietary processed Rx beam information as assistance information from UE to NW, including measured Rx beam information, expected Rx beam information, and best Rx beam information.*  *Proposal 18: Regarding the data collection for AI/ML model training at NW side, study potential specification impact on measurement report:*  *• UE measures the beams of Set A and reports M1 L1-RSRPs optionally with M2 RS indicators, where M1 and M2 can be larger than 4.*  *- If M1 is equal to the number of beams or beam pairs in Set A (noted as X), corresponding RS indicators may be not needed.*  *- If M1 is smaller than X/2, corresponding M2 RS indicators are needed*  *- If M1 is smaller than X, but larger than X/2, RS indicators are needed for indicating M2 beams or beam pairs in Set A not included in the measurement report.*  *Proposal 19: Regarding the data collection for AI/ML model training at NW side, study potential specification impact on report overhead reduction:*  *• Reducing unnecessary L1-RSRP report where the omitted L1-RSRPs may be lower than a pre-defined threshold*  *Proposal 20: Regarding the data collection for AI/ML model training at NW side, study potential specification impact on quantization enhancement for RSRP quality improvement:*  *• High-precision L1-RSRP quantization*  *• Multi-resolution L1-RSRP quantization, e.g. high-resolution quantization for a group of best RSRPs and low-resolution quantization for others* |
| OPPO[6] | *Observation 5: For the simplest case of BM-Case1 and BM-Case2 with NW-side AI/ML model trained at NW, the model inputs (via enhanced beam reporting) and labels (via legacy beam reporting) for training can be reported by UE.*  *Proposal 2: Study data collection for AI/ML model training with enhanced beam reporting mechanism (more than 4 beams per reporting instances) as a starting point.* |
| Spreadtrum[7] | *Proposal 5：For NW-side model, UE report should be enhanced.*  *- the number of bits that can be carried by the UE report should be expanded;*  *- reduce the report overhead by omitting CRIs.*  *Proposal 6：For NW-side model,* *beam management configuration should be enhanced to inform UE the association between configured measurement resources and model input/output.* |
| Nokia[8] | *Proposal 29. For BM-case1 and BM-case2, dedicated RS measurements or reporting framework is not considered for model training.*  *• Note: It is up to the implementation to handle model training*  *Proposal 31. For NW-sided BM-case1 and UE-sided BM-case2, RAN1 shall support the CSI reporting enhancement of reporting more than 4 beams and associated L1-RSRP in a beam report to enable data collection at the NW side.*  *Proposal 32. For NW-sided BM-case1/2, discuss signaling of configuring UE for data recording and reporting for beam measurements of Set B/A corresponding to the failure instances of the NW-sided model.* |
| CATT[9] | *Proposal 2: For DL beam pair prediction with a UE-side model, study the following aspects:*   * *For model training at NW side, study how to report relative Rx beam information when preserving sensitive proprietary information;* * *For model training at UE side, study how to send/report relative Tx beam information when preserving sensitive proprietary information.*   *Proposal 6: Regarding the training data collection for AI/ML model training at NW side, study the following options as a starting point for the contents of collected data:*   * *Opt.1: UE sends M1 L1-RSRPs optionally with the RS/beam (pair) indicators based on the measurement corresponding to a beam set, and optionally sends the beam pattern information (e.g., the Set B beam pattern);* * *Opt.2: UE sends M2 L1-RSRPs optionally with the RS/beam (pair) indicators based on the measurement corresponding to a beam set, sends M3 L1-RSRPs optionally with the RS/beam (pair) indicators based on the measurement corresponding to another beam set, and optionally sends the beam pattern information (e.g., the Set B beam pattern);* * *Opt.3: UE sends M4 L1-RSRPs optionally with the RS/beam (pair) indicators based on the measurement corresponding to a beam set, sends M5 RS/beam (pair) indicators based on the measurement corresponding to another beam set, and optionally sends the beam pattern information (e.g., the Set B beam pattern);* * *FFS: the details of beam pattern information.* |
| Intel[10] | *Proposal 7: For AI/ML model training and inference at NW side, support larger than 4 beams to be reported in one beam reporting instance with potential beam reporting over MAC-CE.* |
| Ericsson[14] | *Observation 1 It is necessary to study the training data collection mechanism for NW-side AI/ML model for the completeness of the AI/ML beam prediction use case*  *Observation 2 An RRC-message based approach is best suitable for training data collection for NW-sided beam prediction*  *Observation 3 The standardized data collection mechanism should enable UE logging and reporting of collected data to the NW*  *Observation 4 For training data collection for NW-sided beam prediction, a measurement occasion configuration needs to be designed to configure/indicate one or multiple sets of SSB/CSI-RS resources for L1-RSRP measurements, where a single set of SSB/CSI-RS resources is associated to a union of the Set A and Set B of beams, as examples shown in Figure 2 and Figure 3.*  *Observation 5 Data collection is seldom performed, and the payload size per UE report for the beam management use cases is not expected to be too large (e.g., around 771 bits for L1-RSRP measurements of 192 CSI-RS beams).*  *Proposal 3 Regarding the training data collection mechanism for NW-side AI/ML model trained at NW side, study the following options as a starting point for the contents of collected data.*  *• Opt.1: UE reports M1 L1-RSRPs (corresponding to M1 beams) corresponding to a beam set (e.g., Set A, Set A+B, Set B), where M1 can be larger than 4. FFS: the range of M1*  *Proposal 4 Regarding the training data collection mechanism for NW-side AI/ML model training trained at NW side, study beam-management-specific potential specification impact (if necessary) from the following additional aspects*  *o Mechanism of the reporting, e.g., RRC signaling, L1 signaling*  * Measurement RS configuration*  * Signaling and/or condition(s) to trigger reporting logged data*  *o Information corresponding to the reported data samples, e.g. non-radio information such as timestamps*  *o Signaling and/or condition(s) to trigger/stop data logging (including buffering)*  * candidate values of data collection duration and measurement occasion intervals.*  * Event types to trigger data logging*  *o Measurement accuracy requirements*  *o Reporting overhead reduction*  * Quantization of the measurement results (e.g., L1-RSRP)*  * Other aspects (e.g. remove duplicated samples, how to collect relevant samples, …)* |
| Fujitsu[15] | *Proposal 1: Regarding the data collection for AI/ML model trained at NW-side, study the potential specification impacts on the UE behavior of beam reporting.*  *Proposal 2: Regarding the data collection for AI/ML model trained at NW-side, considering the non-real time data to be collected, study mechanism on attached timestamp information along with collected data.*   * *FFS: the method to generate the timestamp*   *Proposal 3: Considering the performance of AI/ML model and the overhead of beam reporting, study the necessity and scheme on quantization of beam reporting in the data collection, inference and monitoring on AI/ML model.*  *Proposal 4: Regarding the data collection for AI/ML model training at NW side, the mechanism of reporting overhead reduction is suggested to be studied.*  *Proposal 5: Study the potential specification impacts on the enhanced signaling/procedure of reporting configuration for data collection.* |
| Xiaomi[16] | *Proposal 23: Support following options for the content of the collected data for model training for different set B and set A, and different model output.*   * *Option 1: L1-RSRP of set B and L1-RSRP of set A* * *Option 2: L1-RSRP of set A* * *Option 3: L1-RSRP of set B and Top-K beams ID*   *Proposal 24: Support to define a time window for each report to include more than one data sample and configure a number of report to stop the data collection.* |
| Samsung[19] | *Proposal 1. For BM-Case1 with a network-side AI/ML model, for data collection, study the potential impact of L1 reporting enhancement:*   * *One-shot L1 report with measurement results for one or more beams, where the measurement results of each beam are associated with the corresponding labels (e.g., timestamp, UE speed, SNR, etc.)* * *The handling/buffering for collected data before one-shot L1 reporting* |
| CMCC[22] | *Proposal 1: Regarding the training data collection mechanism for NW-side AI/ML model ~~training~~ trained at NW side, study the following options as a starting point for the contents of collected data*   * *Opt.1: UE sends M1 L1-RSRPs (corresponding to M1 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to a beam set (e.g., Set A, Set A+B), where M1 can be larger than 4*   + *FFS: the range of M1* * *Opt.3: UE sends M4 L1-RSRPs (corresponding to M4 beams) optionally with the indication of beams (beam pairs) ~~indicators~~ based on the measurement corresponding to a beam set (e.g., Set B), sends the indication of M5 beams (beam pair) based on the measurement corresponding to another beam set (e.g., Set A), where M4 can be larger than 4*   + *FFS: the range of M4, M5* * *Other option(s) is not precluded*   *Proposal 2: For Tx-Rx beam pair prediction with AI/ML model training at NW side, study how to define and map the beam pair ID to align the understanding between the NW and the UE.*  *Proposal 3: For DL Tx beam prediction with AI/ML model training at NW side, the Rx beam assumption should be aligned between the network and UE.* |
| MediaTek[23] | *Proposal 1: Support the current shape of Proposal 3.2.1 regarding the contents of collected data for training data collection mechanism for NW-side AI/ML model trained at NW side.*  *Proposal 2: Support the current shape of proposal 3.2.2 regarding studying the necessity and beam-management-specific potential specification impact of the training data collection mechanism for NW-side AI/ML model trained at NW side.* |
| Lenovo[26] | *Proposal 6: Study data collection procedure to support both UE-side and NW-side AI/ML model training and model update*  * For UE-side model training, study procedure to support UE triggered data collection for model update*  * For NW-side model training, support to report larger number of beams in one beam report.*  *○ FFS: Beam report format, e.g., beam report via MAC CE or RRC* |
| NEC[28] | *Proposal 4: Regarding the training data collection mechanism for NW-side AI/ML model training trained at NW side, study unnecessary reporting overhead reduction.* |
|  |  |

###### Proposal 2.2.1

In previous meetings, some alternatives were proposed for discussion with the aim to better understand different proposals from companies. In this meeting, many companies continue the discussions based on Proposal 3.2.1 of the last meeting. Based on the tdocs, the proposal (or some options) seems acceptable to most companies.

Although some companies show their preference on some option(s), the views are still quite diverging and down-selection will need further discussion. Meanwhile, some companies suggest to focus on the detailed spec design. As an SI of AI/ML, it would be beneficial for the group to have solid understanding on the underlaying schemes assumed by different companies for the training data collection. The down-selection and spec design can be discussed later.

Based on the above information, moderator feels the group can continue the discussion based on Proposal 3.2.1 of the last meeting. In previous meetings, some companies commented that there is some overlapping between Option 1 and Option 2. For example, if NW triggers two reporting from UE as in Option 1, it may get the same/similar information for AI/ML training compared to Option 2. In this sense, Option 1 and Option 2 may have similar spec impact and functionality. Thus, the original Option 2 of Proposal 3.2.1 is removed to simplify the proposal. Accordingly, the following proposal is provided for further discussion.

The related proposals in tdocs are as below.

* FUTUREWEI: Proposal 6
* Huawei: Proposal 15
* H3C: Proposal 4
* ZTE: Proposal 13, 16, 17
* vivo: Proposal 18
* OPPO: Proposal 2
* Spreadtrum: Proposal 5
* Nokia: Proposal 31
* CATT: Proposal 6
* Intel: Proposal 10
* Ericsson: Proposal 3
* xiaomi: Proposal 23
* CMCC: Proposal 1
* MediaTek: Proposal 1

***Proposal 2.2.1: Regarding data collection for NW-side AI/ML model, study the following options (including the combination of options) for the contents of collected data***

* ***Opt.1: M1 L1-RSRPs (corresponding to M1 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to a beam set (e.g., Set A, Set A+B, Set B), where M1 can be larger than 4***
  + ***FFS: the range of M1***
* ***Opt.2: M2 beam (beam pair) indices based on the measurement corresponding to a beam set (e.g., Set A), where M2 can be larger than 4***
  + ***FFS: the range of M2***
  + ***FFS: How to select the M2 beam (e.g. M2 strongest beam(s))***
* ***Note1: Data collection for model training may not have 3GPP specification impact***
* ***Note2: Overhead, UE complexity and power consumption should be considered for the above options***

|  |  |
| --- | --- |
| Company | Comments |
| Nokia/NSB | First, we are not clear on the need of this proposal. As discussed in the last meeting, the data collection shall mainly focus the performance monitoring or model fine-tuning, and focus shall not be model training.  Third, just to make things clear, we should be fine with something like below,  Mod: Understood your intention, but as you can see, many companies are still considering some mechanism(s) for training data collection.  ***Proposal 2.2.1: Regarding the data collection mechanism for NW-side AI/ML model, study the following for the contents of collected data***   * ***Option 1: M1 best beam indices and corresponding L1-RSRPs based on the measurement corresponding to a beam set where M1 can be larger than 4***   + ***FFS: the range of M1***   + ***Beam set can be fully transparent to the UE (e.g., exact knowledge about Set A/B are not known to the UE)*** * ***Option 2: M1 best beam indices and corresponding L1-RSRPs based on the measurement a first beam set (e.g., Set B), M2 best beam indices based on the measurement corresponding to a second beam set (e.g., Set A), where M2 can be larger than 4***   + ***FFS: the range of M1, M2***   + ***Beam sets are partly transparent to the UE (e.g., there may be some knowledge about Set A/B to the UE)***   Mod: Most modifications are reflected in the new version. The two newly-added parts is not included. Regarding the “beam indices”, please see vivo/ZTE’s comment |
| Google | We suggest adding the following options on top of Nokia’s version. Since this is for NW model, there is no need for UE to select the M1 beams, but the NW can configure the M1 beams as CMR. The UE just reports the L1-RSRP for all the configured beams.   * ***Option 3: A first report with L1-RSRPs based on the measurement corresponding to all the configured M1 beams in a beam set where M1 can be larger than 4 (Set B), and a second report with M2 beam indexes for the beams with highest L1-RSRP in a beam set with M3 configured CMRs (Set A)***   + ***FFS: the range of M1, M2, M3***   + ***FFS: Whether the first report and second report are in a single report or separate reports***   Mod: The current options are inclusive and doesn’t say how to select/decide the reported beams. The further details (e.g., how to decide the reported beams) can be discussed later. |
| New H3C | We are ok in general, for opt.2, maybe some typos:   * ***Opt.2: UE sends M2 L1-RSRPs (corresponding to ~~M4~~M2 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to a beam set (e.g., Set B), sends M3 beams (beam pair) based on the measurement corresponding to another beam set (e.g., Set A), where M2 can be larger than 4***   Mod: fixed |
| NTT DOCOMO | Support the updated version from Nokia. |
| LG | To us, the difference between option1 and option2 is not clear. If there is no inter-relation between the first report and the second report, option2 seems a special configuration with option1.  Mod: Note1 is removed. For option 1, the reporting will always include the L1-RSRP. But for Option 2, only beams (without L1-RSRP) are included for the 2nd set |
| vivo | OK with the current formulation of the proposal. We don’t support the version from Nokia. In either option, the reported beam indicators do not need to correspond to the reported RSRPs. Actually, if the reported number of RSRPs need to be large for training, it is more efficient from overhead perspective if the reported beam indicators are the ones corresponding to the RSRPs which are not reported. Further, if the all the RSRPs in set A are reported for training, there is no need to report any beam indicators. |
| HW/HiSi | We are supportive to study the spec impact for data collection in general.  As pointed out by Nokia, that could (or should) be applicable to both training and but also monitoring and fine-tuning. We therefore agree with Nokia to remove the word “training” from the main bullet.  Can the intention of Note2 be clarified? Does it mean that the purpose (e.g. training) is transparent, or does it mean that Option 1 or Option 2 may not be needed at all and legacy handling could be used instead?  **In principle, with a clarification on Note and with removing “training” from the main bullet we could be fine.** |
| Spreadtrum | We have some confusion about option 2. Does it mean that UE will send two reports to the gNB? If not, we agree with LG, option2 is a special case of option 1 actually.  Mod: Note1 is removed. |
| CAICT | We are fine with the proposal and update with removing “training” from the main bullet.  Mod: Updated. |
| CATT | As discussed in our contribution, it is possible that NW can configure multiple Set B patterns for UE measurement, and UE reports the best Set B patterns to NW. Therefore, the pattern information can also be reported by UE. We suggest the following update:  ***Proposal 2.2.1: Regarding the training data collection mechanism for NW-side AI/ML model trained at NW side, study the following options as a starting point for the contents of collected data***   * ***Opt.1: UE sends M1 L1-RSRPs (corresponding to M1 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to a beam set (e.g., Set A, Set A+B, Set B),*** ***and optionally sends the beam pattern information (e.g., the Set B beam pattern), where M1 can be larger than 4***   + ***FFS: the range of M1*** * ***Opt.2: UE sends M2 L1-RSRPs (corresponding to M4 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to a beam set (e.g., Set B), sends M3 beams (beam pair) based on the measurement corresponding to another beam set (e.g., Set A), and optionally sends the beam pattern information (e.g., the Set B beam pattern), where M2 can be larger than 4***   + ***FFS: the range of M2, M3***   + ***Note1: From UE perspective, the measurement and reporting related to one beam set may be separate from/transparent to the operations related to another beam set*** * ***Note2: Data collection for model training may be implemented by gNB in a transparent way***   ***Note3: Potential down-selection/prioritization will be discussed later***  Mod: “indication of beam (beam pairs)” is quite inclusive. It is moderator’s understanding that the new part can be regarded as a special type of “indication of beam”. |
| CMCC | Ok in general. In option 1, the motivation to measure only Set B is not clear to us.  Mod: For option 1, gNB may trigger two independent reporting (from UE perspective), one is for Set A and one is for Set B. Then, NW will associate the input and labels based on the two reporting from UE |
| Ericsson | Prefer the update from Nokia. |
| Xiaomi | We are general fine with the FL’s proposal. and we prefer to remove the “set B” in “……***to a beam set (e.g., Set A, Set A+B, ~~Set B~~), where M1 can be larger than 4***” since it doesn’t work for set B.  Mod: Please see the reply to CMCC |
| Samsung | We generally fine with the proposal from FL. There is no need to differentiate the purpose of data collection (e.g., for training or fine-tune). Hence, suggest to remove ‘train’ in the main bullet as below.  ***Proposal 2.2.1: Regarding the ~~training~~ data collection mechanism for NW-side AI/ML model ~~trained at NW side~~, study the following options as a starting point for the contents of collected data***  Mod: Updated |
| Fujitsu | Generally, support this proposal and it’s better to remove the “training” from main bullet.  Mod: Updated |
| NVIDIA | Support the proposal in principle. |
| Qualcomm | At a high level, we share the same thought as Nokia that the need for this proposal is not clear. In the proposal text, we talk about “contents of collected data”, so in our view, we should avoid mentioning “UE sends M1 L1-RSRPs”, which is addressed in Nokia’s update. Also, for some reason Note 4 from the previous version was removed which is added back in the suggested wording. Following up on Nokia’s suggested update, we suggest the following update:  ***Updated Proposal 2.2.1: Regarding the ~~training~~ data collection mechanism for NW-side AI/ML model, study the following for the contents of collected data***   * ***Option 1: M1 best beam indices and corresponding L1-RSRPs based on the measurement corresponding to a beam set where M1 can be larger than 4***   + ***FFS: the range of M1***   + ***Beam set can be fully transparent to the UE (e.g., exact knowledge about Set A/B are not known to the UE)*** * ***Option 2: M1 best beam indices and corresponding L1-RSRPs based on the measurement corresponding to a first beam set (e.g., Set B), M2 best beam indices based on the measurement corresponding to a second beam set (e.g., Set A), where M2 can be larger than 4***   + ***FFS: the range of M1, M2***   + ***Beam sets are partly transparent to the UE (e.g., there may be some knowledge about Set A/B to the UE)*** * ***Note2: Data collection for model training may be implemented by gNB in a transparent way*** * ***Note3: Potential down-selection/prioritization will be discussed later*** * ***Note4: Overhead, UE complexity and power consumption should be considered for the above options***   Mod: Note 4 is added back |
| Lenovo | If this proposal is for initial model training, option 1 may be enough at least for the case the Set B is a subset of Set A since the pattern of Set B may de determined based on the model training. Option 2 is more applicable for model update.  We can support Nokia’s version and we want to clarify that M1 and M2 can be the number of all the beams configured in Set A or Set B.  Mod: “training” is removed. Regarding the clarification, it is next-level detail. We can discuss it latter. |
| Apple | Is Option 2 mainly about the case where set B is not a sub-case of Set A (e.g., Set B is for SSB beams, and Set A is or CSI-RS beams)?  Mod: If moderator understood correctly, it can also be used for the case where set B is a subset of Set A. The proponents of option 2 can make further clarification. |
| MediaTek | We are fine with this proposal, only one typo of Opt.2:  ***Opt.2: UE sends M2 L1-RSRPs (corresponding to M4 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to a beam set (e.g., Set B), sends M3 beams (beam pair) indices based on the measurement corresponding to another beam set (e.g., Set A), where M2 can be larger than 4***  Mod: updated |
| ZTE | General fine with the proposal from FL. Besides, we agree with vivo that the reported RSRP does not always need to be associated with a beam indicator. For example, if all measurement results of set A/B need to be reported, the beam indicator can be obtained implicitly from the reporting order of all measured RSRPs, resulting in a significant reporting overhead reduction. |
| Futurewei | We are OK with the proposal (with the typo *M4* corrected, should be *M2*).  Mod: Fixed |
| Mod | The proposal is updated based on the comments/suggestions. (U1) |
| NEC | Support the updated version in general. But we would like to clarify the consideration of Note4. Is it for further down-select from Opt.1and Opt.2?  Mod: Do you refer to the last note? Some companies suggested to consider the aspects of Overhead, UE complexity and power consumption. In my understanding, this note is to reminder the group to consider these aspects. Whether there will be any down-selection may depend on more aspects. |
| Samsung2 | In terms of the updated proposal 2.2.1, we think that L1-RSRPs of M3 beams are also useful for the purpose of data collection.  Hence, the following change is suggested:   * ***Opt.2: ~~UE sends~~ M2 L1-RSRPs (corresponding to ~~M4~~ M2 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to a beam set (e.g., Set B), ~~sends~~ and M3 beam~~s~~ (beam pair) indices and optionally with the corresponding L1-RSRPs based on the measurement corresponding to another beam set (e.g., Set A), where M2 can be larger than 4***   Mod: Share similar view as Huawei (please see it below). The option2 suggested by SS can be implemented by trigger two reporting of opiton1 |
| Hw/HiSi | We are ok in principle with the updated proposal from the FL, but have one question for our understanding on Opt1: If Set B is different from Set A, would Opt.1 then still be possible to use?  Mod: My understanding is Yes. gNB can configure a “super set” consisting of Set A and Set B  @SS, FL: Regarding the suggested update in blur from SS, could you please give an example when it would be useful to also report optionally the L1-RSRP for e.g. Set A? In my understanding, the beam indicator from Set A would be used to label the best beam, and in this case RSRPs would not be needed- |
| InterDigital | It seems that our previous input was ignored due to NVIDIA’s input. I am not sure some company ignores checkout then why this mechanism is needed. We copy the previous comment here.  As this proposal is for NW side model, we are not sure that how the UE can report not measured beams in Option 2. Before agreeing this proposal, necessity of Option 2 for NW side model should be clarified.  Mod: Option 2 can include this case:   * gNB configures two reporting (namely A and B) * For reporting A, L1\_RSRP + Beam ID is reported * For reporting B, Top-K beam is reported (No L1-RSRP)   If Option 1 is used, L1\_RSRP should be reporting for both A and B, which may lead to more overhead.  If we need to agree this proposal, then we prefer Nokia’s update.  Mod: The current version is based on Nokia’s suggestion by considering other companies’ input  For Option 3 proposed by Google, we think that it’s a joint operation of Option 1 and Option 2. In that regard, we don’t need to explicitly mention it.  Mod: Not included in current version. |
| Futurewei | OK with the proposal. |
| Qualcomm | First, we are not sure why the Notes 1 to 3 were removed.  Mod: In Nokia’s version, Note1 is removed and some new sub-bullets is added. In my understanding the new sub-bullet may not be always the case. Please see my reply to Nokia  Note2 was suggested by Nokia in previous meeting. Now their new version doesn’t include it. Thus, note2 is removed  It seems Note3 doesn’t provide much information. Since the proposal is suggest to study some options, the potential down-selection/prioritization is always possible  If companies think it is better to add back Note1/2/3, we can add back them.  Second, our update above which is a follow-up on the update from Nokia makes the distinction between Option 1 and Option 2 a bit clearer. With the current wording, and with the notes removed, the intention of Option 2 and it’s distinction with Option 1 is not clear.  Mod: For option 1, a reporting will always include L1-RSRP. For Option 2, a reporting may only include beam ID. Thus, they are different. Please see the reply to IDC as well |
| ZTE | Fine with the updated proposal from FL. |
| Lenovo | Fine with the updated proposal |
| Xiaomi | As we commented before, It is talking about the content of collected data, not the content in each report. So it is not possible that the content of collected data only contains the M1 L1-RSRPs of set B. If it is allowed, Opt .2 should be updated as below, i.e., to change ‘and’ to ‘and/or’ since each part can be a measurement report.   * ***Opt.2: ~~UE sends~~ M2 L1-RSRPs (corresponding to ~~M4~~ M2 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to a beam set (e.g., Set B), ~~sends~~ and/or M3 beam~~s~~ (beam pair) indices based on the measurement corresponding to another beam set (e.g., Set A), where M2 can be larger than 4***   Mod: updated |
| Samsung3 | @HW: with the blue part added in my last comment, data collection for the case of Set A and Set B are different can be resolved.  Mod: It can also be solved by Option 1.   * gNB configures two reporting (namely A and B) * reporting A is for Set A * reporting B is for Set B * For UE perspective, these two reporting can be configured independently. |
| Mod | @Nokia, LGE, QC, IDC:  Difference between Option 1/2: For option 1, a reporting will always include L1-RSRP. For Option 2, a reporting may only include beam ID. Thus, they are different.  @Nokia, QC:  Relationship between sets: For option 2, it is still possible beam sets are transparent to the UE. Option 2 can include this case:   * gNB configures two reporting (namely A and B) * For reporting A, L1\_RSRP + Beam ID is reported * For reporting B, Top-K beam is reported (No L1-RSRP) * For UE perspective, the two different reporting may be configured independently   If Option 1 is used, L1\_RSRP should be reporting for both A and B, which may lead to more overhead.  In the current version of Proposal 2.2.1, the changes not getting any comment are un-highlighted. Only the new modifications or the modifications getting comment are highlighted by RED. The same principle is also applied for other proposals |
| ZTE | Support. |
| Ericsson | Our understanding is same as FL clarification above. And this explanation is much clearer than the updated proposal.   ***Proposal 2.2.1: Regarding data collection for NW-side AI/ML model, study the following options for the contents within one reporting instance ~~of collected data~~***   * ***Option 1: L1-RSRP measurement report on a single beam set (e.g. set A+B)*** * ***Option 2: Measurement report on two beam sets***   + ***L1-RSRP report on a first beam set (e.g. set B)***   + ***Beam indications (e.g. strongest beam) of a second beam (e.g. set A)*** * ***FFS: Number of reported beams in each option*** * ***Note: If Option 1 is used, it includes reporting for both A and B, which may lead to more overhead.***   Mod: If the data collection is based on L1 signaling, the above version seems better. Meanwhile, RAN2 are study the framework of data collection. Thus, our proposal should be also applicable to the framework based on higher layer signaling (if agreed by RAN2). For example, NW may configure UE to report the result with a duration, then there will be multiple of training data. In this case, not sure whether “reporting instance” is clear or not. Additional comments   * Indications of the beams (from UE to NW) may be needed in some cases as suggested by some companies * Reporting of up to N (N>4) beams is one of the key aspects |
| Xiaomi | Support |
| HW/HiSi | Support |
| Lenovo | Thanks FL’s interpretation on option 1 and option 2. But the difference between option 1 and option 2 as Mod explained is not reflected in the current wording.  I understand the data collection in this proposal is based on Rel-17 beam reporting framework, i.e., the beam report is carried by UCI. M1 beams in option 1 and M2 beam and M3 beams in option2 are reported in different beam report, and each of M1, M2, M3 may be larger than 4.  Mod: The way based on Rel-17 beam reporting framework is possible. Meanwhile, RAN2 are studying the framework of data collection based on higher layer signaling. Thus, our proposal is intended to be applicable to all possible solutions.  On the example provided in option 1, I fail to see the case for M1 beam reporting based on set A+B, a bit more clarification is highly appreciated.  Mod: One example is as below:   * NW can configuration a set (consisting of all beams from A+B) for measurement * UE will report all the L1-RSRP (+ beam IDs). * NW can get the training data including the model inputs and the corresponding labels |
| Qualcomm | Thanks, FL, for the clarification above which was helpful.  When we talk about the *contents* of the collected data in the proposal, what is the intention of *Set A+B*? We may collect the measurements (L1-RSRP + beam ID) associated with Set A and/or Set B, but what does Set A+B mean for *collected* data. Additionally, why should we repeat Opt.1 in Opt. 2, when we are talking about *collected* data?  When we study the *contents* of the collected data, we should avoid discussions such as *whether Set A+Set B is reported in a single reporting instance*, etc. What matters in the context of this proposal is what type of data is collected, reporting is a separate discussion.  Based on the above clarifications by FL, we propose a simplified version which attempts to capture the intention of the proposal in a concise and succinct manner below:  ***Updated Proposal 2.2.1: Regarding data collection for NW-side AI/ML model, study the following options for the contents of collected data, based on the measurement corresponding to a beam set***   * ***Opt.1: M1 L1-RSRPs (corresponding to M1 beams) optionally with the indication of beams (beam pairs), where M1 can be larger than 4 (e.g., for Set B, Set A)***   + ***FFS: the range of M1*** * ***Opt.2: M2 beam indices, were M2 can be larger than 4 (e.g., for Set A)***   + ***FFS: the range of M2*** * ***Note 1: Data collection may be implemented by gNB in a transparent way*** * ***Note 2: Overhead, UE complexity and power consumption should be considered for the above options*** |
| Mod | The proposal seems clearer if we follow the direction of Ericsson/QC. Thus, the proposal is updated accordingly   * “***including the combination of options***” is added in the main bullet to cover the Option 2 in Ericsson’s version * In opt.2, only keep the contents of beam indices * Some note is added back |
| NEC | Mod: Do you refer to the last note? Some companies suggested to consider the aspects of Overhead, UE complexity and power consumption. In my understanding, this note is to reminder the group to consider these aspects. Whether there will be any down-selection may depend on more aspects.  Yes. Thanks for FL’s clarification. We support the updated proposal. |
| Samsung3 | Pretty much agree with QC. The key of data collection of NW model is what kind of information is needed by NW.  For regression model, L1-RSRP is needed for model input as well as for model output (i.e., labeling in the training or fine-tuning phase). For classification model, L1-RSRP is needed for model input; while best beam index and optionally L1-RSRP are needed for model output (i.e., labeling in the training or fine-tuning phase).  Basically, L1-RSRP and beam indices are the essence of data collection which is well reflected by the QC’s update.  How the above information is conveyed (for example, via in one or more report instance) to gNB is up to the discussion in WI phase.  We think it is already agreed to study beam report with more than 4 measurement results. There is no need to repeat the restriction here. Hence, we have the following suggestion.  ***Updated Proposal 2.2.1: Regarding data collection for NW-side AI/ML model, study the following options for the contents of collected data, based on the measurement corresponding to a beam set***   * ***Opt.1: M1 L1-RSRPs (corresponding to M1 beams) optionally with the indication of beams (beam pairs)~~, where M1 can be larger than 4~~ (e.g., for Set B, Set A)***   + ***FFS: the range of M1*** * ***Opt.2: M2 beam indices~~, were M2 can be larger than 4~~ (e.g., for Set A)***   + ***FFS: the range of M2***   Mod: Our previous agreement is only for model inference. This proposal may be applicable to initial training data, fine-tuning, …. Thus, there are still needed |
| Ericsson | Share the view by Samsung. For option 2, it is not clear on how the beam indices are derived. We assume the common understanding is that M2 strongest beams are reported. However, there might be other examples on how to derive the beam indices (e.g. beam indices within 10 dB of the strongest beam are reported). We prefer companies to provide their views on this, hence, FFS is added below.  We are not sure what is meant by “gNB transparent way”. Suggest the amendments below.  ***……***   * ***Opt.2: M2 beam indices~~, were M2 can be larger than 4~~ (e.g., for Set A)***   + ***FFS: the range of M2***   + ***FFS: Methods to derive M2 beam indices (e.g. M2 strongest beam(s))*** * ***Note1: The options may not have 3GPP specification impact ~~be implemented by gNB in a transparent way~~***   Mod: For the first comment, please see the reply to SS. The two modifications are updated with some rewording |
| HW/HiSi | Ok, with the understanding that Opt1 works for both Set B is the subset of Set A and different from Set A. |
| Mod | The proposal is updated based on two modification suggested by Ericsson. |
| New H3C | OK |
| Spreadtrum | OK |
| Futurewei | Support |
| ZTE | Support |
| LG | On the second FFS of Opt2, why beam selection method is not needed to be studied for Opt1? Seems some alignment is needed across Opt1 and Opt2. |
| CMCC | Ok. |
| HW/HiSi | Ok. |

###### Proposal 2.2.2

In previous meetings, we agreed to study the following aspect

• Signaling/configuration/measurement/report for data collection, e.g., signaling aspects related to assistance information (if supported), Reference signals

In the last meeting, Proposal 3.2.2 was suggested to study these additional aspects, but no consensus was made. By reading the tdocs, Proposal 3.2.2 of the last meeting can still cover most of the potential spec impacts suggested by companies. As a study item for AI/ML, it would be good to have a record to capture a list of all the potential aspect that the group have discussed and studied, and to encourage companies to think about more aspects. Thus, we can continue the discussion based on Proposal 3.2.2 of the last meeting. Accordingly, the following proposal is provided for further discussion.

* The proposal keeps “if necessary” and we may make down-selection based on further study/discussion.
* Some modifications are made with two intentions: 1. To address the concerns raised in the last meeting; 2. To reflect some proposals suggested by tdocs.
  + “user plane, control plane” is changed to “other higher-layer mechanism”
  + Note2 is added
  + “data quality” is added as an example in the 2nd bullet

The related proposals in tdocs are as below:

* Huawei: Proposal 17, 18, 21
* H3C: Proposal 3, 4
* vivo: Proposal 19, 20
* Ericsson: Proposal 4
* Fujitsu: Proposal 2, 3,4,5
* xiaomi: Proposal 24
* Samsung: Proposal 19
* MediaTek: Proposal 2
* Lenovo: Proposal 6
* NEC: Proposal 4

***Proposal 2.2.2: Regarding data collection for NW-side AI/ML model, study necessity, benefits and beam-management-specific potential specification impact from RAN1 point of view on the following additional aspects***

* ***Mechanism related to the reporting***
* ***Additional information for content of the reporting***
  + ***FFS: Information associated with or configured for the reported data samples, e.g., timestamps, SNR, data quality, etc.***
* ***Reporting overhead reduction***
* ***Note1: non-3GPP based solution is a separate issue.***
* ***Note2: The framework corresponding to higher layer(s) are up to the associated WG(s)***
* ***Note 3: Overhead, UE complexity and power consumption should be considered***

|  |  |
| --- | --- |
| Company | Comments |
| Nokia/NSB | Assuming that RAN2 defines a framework for data collection for model training, RAN1 focus shall only be the content of the data collection.  For performance monitoring and fine-tuning, RAN1 can discuss measurement and reporting enhancements. But, as we have some agreements on performance monitoring more or less covering the bullets above, we do not think this proposal is needed.  Mod: During the discussion in the RAN plenary, it was clarified that one WG is doing something can not be an excuse to stop other WG(s) to do some related study. In moderator’s understanding, it is also applicable for this case. In this proposal, each sub-bullet may have RAN1 impact:   * 1st bullet: some companies proposed to use L1 signaling for training data collection * 2nd bullet: contents that will reported from UE to NW * 3rd bullet: If there is L1 signaling for data collection, the triggering/stop may have RAN1 spec impact * …   In order to emphasize that, “RAN1” is added in the main bullet. |
| Google | Agree with Nokia  Mod: Please see the reply to Nokia |
| New H3C | We are ok. |
| LG | Agree with Nokia  Mod: Please see the reply to Nokia |
| vivo | OK with the current formulation of the proposal. |
| HW/HiSi | Support the main bullet in principle.  We think using the word “training” could be removed from the main bullet since it might cause confusions at this stage (is it for initial training, or training for model updating/fine-tuning or maybe even for monitoring)? In principle, the bullets listed in the proposal could be studied for initial training, fine-tuning/model updating, monitoring.  For the sub-bullets, the current proposal suggest that the attributes are associated with the training data and sent to the NW. This can cost quite some overhead in reporting. A more efficient way for at least some characteristics (like the required data quality) would be to let the NW configure them. We are therefore making the following suggestion for modification:  In our opinion the UE speed is as privacy information (similar to UE moving direction or UE location), also it is possible to obtain the UE trajectory from the speed. Therefore, if UE speed is intended to be included in the UE report, then we should ask RAN2/SA3 to look into the feasibility and how to do it. We are suggesting the following Note 3  ***Updated Proposal 2.2.2: Regarding the ~~training~~ data collection mechanism for NW-side AI/ML model trained at NW side, study necessity and beam-management-specific potential specification impact (if necessary) from the following additional aspects***   * ***Mechanism of the reporting, e.g., RRC signaling, L1 signaling, other higher-layer mechanism*** * ***Information associated with or configured for the reported data samples, e.g., timestamps, UE speed, SNR, data quality, etc.*** * ***Signaling and/or condition(s) to trigger/stop data logging (including buffering) and/or reporting*** * ***Quantization of the measurement results (e.g., L1-RSRP)*** * ***Reporting overhead reduction*** * ***Note1: non-3GPP based solution is a separate issue.*** * ***Note2: The detailed designs corresponding to higher layer(s) are up to the associated WG(s)*** * ***Note3: Whether/how to support the report of UE speed needs to be discussed at RAN2/SA3***   Mod: Some modifications are captured in the revision. For UE speed, if some companies think it is also a kind of privacy information, we can discuss it separately. Thus, “speed” is removed |
| Spreadtrum | Agree with the proposal |
| CAICT | Support the proposal and fine with HW’s update. |
| CATT | Support FL proposal. |
| CMCC | Support. |
| Ericsson | Support the update from HW to remove “training”. We agree with Nokia that it could be a RAN2 topic, however some sub-bullets such as the reporting overhead reduction and quantization could be studied in RAN1. Prefer to keep the proposal |
| Xiaomi | Support the proposal and fine with the note 3 added by HW. |
| Samsung | In our view, there is no need to differentiate the purpose of data collection (e.g., for training or fine-tune). Hence, suggest to remove ‘train’ in the main bullet as below.  ***Proposal 2.2.2: Regarding the ~~training~~ data collection mechanism for NW-side AI/ML model ~~trained at NW side~~, study necessity and beam-management-specific potential specification impact (if necessary) from the following additional aspects***  Mod: The modifications are reflected in the revision |
| Fujitsu | Support this proposal but it’s better to remove the “training” in main bullet.  Mod: updated |
| NVIDIA | Support the proposal in principle. |
| Qualcomm | Agree with Nokia  Mod: Please see the reply to Nokia |
| Lenovo | Share similar view with Nokia.  Mod: Please see the reply to Nokia |
| MediaTek | Support |
| Mod | The proposal is updated based on the inputs  @Nokia, Google, LGE, QC, Lenovo: “RAN1” is added in the main bullet to emphasize the scope. Hope it is acceptable |
| ZTE | Support to remove ‘training’ in the main bullet.  Additionally, we suggest to add some examples for the reporting overhead reduction, such as UE-side data filtering. Due to the relatively long data collection duration, the wireless link between gNB and UE may change rapidly, resulting in unguaranteed data sample quality. For example, if the link failure happens due to beam blockage, the beam quality of the measured data sample would be very weak and impacted most by the noise. Thus, such non-valid measured data is useless for the model monitoring or fine-tuning and shall be filter out.  ***Reporting overhead reduction, such as UE-side data filtering***  Mod: The details we can discussed later. Before the discussion, the group may want to better understand what the data filtering is so that we can know whether it have RAN1 or/and RAN2 spec impact. |
| Ericsson | Ok with the update. However, inline with ZTE, we prefer to provide some examples for reporting overhead reduction, where we think Quantization is one example, since higher resolution is not proved to improve prediction performance in 9.2.3.1   * ***~~Quantization of the measurement results (e.g., L1-RSRP)~~*** * ***Reporting overhead reduction (e.g Quantization, UE-side data filtering, avoid reporting duplicated samples, …. )***   Mod: Regarding the examples, please see the reply to ZTE. For a reporting, the spec should specify which quantization mechanism is used. Even if the existing L1-RSRP quantization is reused, we still need an agreement/conclusion. The current wording “quantization of the measurement results” is quite inclusive and doesn’t imply any specific mechanism. |
| NEC | Support the updated version in general. And we are open to add some examples (e.g., UE-side data filtering, avoid reporting duplicated samples) into Reporting overhead reduction. |
| HW/HiSi | Support the updated version.  At this stage, for the sake of progress I think we should avoid adding examples. |
| InterDigital | We believe that this proposal is needed as the proposal includes required details such as quantization and reporting overhead reduction.  As we commented, explicitly mentioning RAN1 specification does not make sense to us as RAN1 features generally require other specification enhancements as well.  In that sense, we propose the following update on the main bullet.  ***Regarding ~~the training~~ data collection ~~mechanism~~ for NW-side AI/ML model ~~trained at NW side~~, study necessity and beam-management-specific potential ~~RAN1~~ specification impact (if necessary) from RAN1 point of view on the following additional aspects***  Mod: Updated |
| Futurewei | OK with the updated version. |
| Qualcomm | We reiterate our earlier position that this proposal is not needed, for the reasons mentioned earlier.  Mod: In the previous round, QC just said “Agree with Nokia”. I have made some reply to Nokia’s comment. Thus, would you like to elaborate a bit more on the technical reasons/concerns? |
| ZTE | We agree with HW that some characteristics (like the required data quality) can be configured by NW. Then, the measured data that cannot meet the configured data quality requirement shall not be reported from reporting overhead reduction perspective. Since the ‘necessity’ to study the additional aspects was already mentioned in the main bullet, we prefer to add some typical examples for the reporting overhead reduction to trigger some further study. The updated version from Ericsson is OK to us. |
| Lenovo | Fine with the update. |
| Xiaomi | Support the updated proposal |
| LG | Even if RAN1 is included in the main bullet, many aspects overlapped with RAN2 discussion, for example, the underlined parts:   * Mechanism of the reporting, e.g., RRC signaling, L1 signaling, other higher-layer mechanism in the first bullet * Signaling and/or condition(s) to trigger/stop data logging (including buffering) and/or reporting).   We may be ok if the proposal is limited to RAN1 feature based solution only, i.e. UCI based beam reporting enhancement. Otherwise, there will be significant overlap with RAN2 discussion.  Mod: Even if RAN2 lead the work and the framework is based on higher-layer signaling, there will be some RAN1 work needed. The proposal is updated according to IDC’s suggestion and it to emphasize the study is only from RAN1 perspective. We will avoid the potential collision between other WG(s)  Note2 is also modified. Hope it can address the concern in some sense. |
| Mod | In the main bullet, “if necessary” is removed since “necessity” has been included, and the “benefits” is also included.  The 4th sub-bullet is removed. There is an agreement for the quantization in Feifei’s session. Thus, there is no much motivation to introduce finer granularity quantization. For the quantization with less bits can be considered under the sub-bullet “reporting overhead reduction” |
| ZTE | We think the measured data that cannot meet the data quality requirement shall not be reported from reporting overhead reduction perspective. Besides, as FL also consider the quantization as an sub-bullet of reporting overhead reduction, it would be better to clarify it and include some typical examples for reporting overhead reduction in the proposal for further study, at least for studying their necessity and benefit.   * ***Reporting overhead reduction, e.g., quantization, UE-side data filtering*** |
| Ericsson | Support |
| Xiaomi | Support the updated proposal |
| HW/HiSi | Support the proposal.  Regarding the comment from ZTE on overhead reduction, we share “*We think the measured data that cannot meet the data quality requirement shall not be reported from reporting overhead reduction perspectiv*e”, but we think we should avoid mentioning examples at this stage. There are many candidates what to do for overhead reduction and a discussion about the details (that many companies than probably want to add) would risk to stall the progress. |
| Lenovo | One comment on the second bullet: what the meaning of ***data quality***? |
| Fujitsu | support |
| Nokia/NSB | We assume that the proposal is generic for data collection, and mainly for model updates or model monitoring. We have the following comments.   * “Mechanism of the reporting”: is this measurement reporting or some other reporting? RAN1 does not have the scope on examples mentioned in the bullet and RAN2 is already in the discussion on the methods of data collection. Suggest only to consider measurement reports the method of reporting, which seems to be fine for fine-tuning or monitoring of a NW-sided model.   Mod: Some companies propose to use L1 signaling (e.g., UCI) for data collection. Thus, the 1st bullet is here   * Timestamps, data quality, etc., are not up to RAN1 to decide.   Mod: Not exactly. These topics are discussed and decided in RAN1 for some 3GPP projects   * Third bullet is also not clear. Can we get an example of signaling impact on RAN1 specs for data logging.   Mod: Some companies propose to use L1 signaling. Thus, there may be some RAN1 work.   * Fourth bullet is OK as long as we discuss reporting RS measurements. * Notes are not related to the RAN1 discussion.   Mod: These notes are added to restrict the scope of this proposal. Let’s here more views.  We suggest the following to keep the scope in RAN1 level.  ***Proposal 2.2.2: Regarding data collection for NW-side AI/ML model, study ~~necessity, benefits and~~ beam-management-specific potential RAN1 specification impact ~~(if necessary)~~ ~~from RAN1 point of view~~ on the following additional aspects***   * ***Mechanism of the RS measurements and reporting~~, e.g.,~~ ~~RRC signaling, L1 signaling, other higher-layer mechanism~~*** * ***Content of the reporting ~~Information associated with or configured for the reported data samples,~~ ~~e.g., timestamps, SNR, data quality, etc~~.*** * ***~~Signaling and/or condition(s) to trigger/stop data logging (including buffering) and/or reporting~~*** * ***~~Quantization of the measurement results (e.g., L1-RSRP)~~*** * ***Reporting overhead reduction*** * ***~~Note1: non-3GPP based solution is a separate issue.~~*** * ***~~Note2: The detailed designs/framework corresponding to higher layer(s) are up to the associated WG(s)~~***   Mod: The 1st/2nd bullet of your version seems not adding much new information, since we have the following agreement  Agreement  For the data collection for AI/ML model training (if supported), study the following aspects as a starting point for potential necessary specification impact:   * Signaling/configuration/measurement/report for data collection, e.g., signaling aspects related to assistance information (if supported), Reference signals * Content/type of the collected data * Other aspect(s) is not precluded |
| Qualcomm | Following up on Nokia’s update, we suggest the following updates:  ***Proposal 2.2.2: Regarding data collection for NW-side AI/ML model, study necessity, benefits and beam-management-specific potential RAN1 specification impact (if any) ~~(if necessary)~~ ~~from RAN1 point of view~~ on the following additional aspects***   * ***Mechanism of the ~~RS measurements and~~ reporting~~, e.g.,~~ ~~RRC signaling, L1 signaling, other higher-layer mechanism~~*** * ***Content of the reporting ~~Information associated with or configured for the reported data samples,~~ ~~e.g., timestamps, SNR, data quality, etc~~.*** * ***~~Signaling and/or condition(s) to trigger/stop data logging (including buffering) and/or reporting~~*** * ***~~Quantization of the measurement results (e.g., L1-RSRP)~~*** * ***Reporting overhead reduction*** * ***Note 1: non-3GPP based solution is a separate issue.*** * ***Note 2: Data collection may be implemented by gNB in a transparent way*** * ***Note 3: Overhead, UE complexity and power consumption should be considered***   ***~~Note2: The detailed designs/framework corresponding to higher layer(s) are up to the associated WG(s)~~*** |
| Futurewei | Support in general. We are not clear on what the “design” means in Note 2.   * ***Note2: The ~~detailed~~ designs/framework corresponding to higher layer(s) are up to the associated WG(s)***   We understand the “framework” could be the data collection framework that RAN2 is discussing. If “design” has no specific meaning, it can be removed.  Mod: “design” is deleted |
| Mod | The proposal is updated based on the comments |
| LG | Similar view with Nokia and Qualcomm. On the updated proposal, suggest to remove the third bullet. |
| Samsung | Generally fine with the update from QC. One wording suggestion is to change ‘Mechanism of reporting’ to ‘Mechanism **related to** reporting’.  Mod: updated |
| Ericsson | Support Also ok with the update from QC. Except their note on “gNB in a transparent way”, since the main proposal outlines the study of potential spec impact, it is not needed. |
| HW/HiSi | Support.  Agree with Ericsson on their comment for Note 2. |
| Mod | The proposal is updated:   * Since several companies suggest to remove the 3rd bullet, it is deleted * One minor wording change based on SS’s comment |
| New H3C | OK |
| Spreadtrum | Generally fine with the updated proposal from FL. Since we have limited this proposal in the main bullet to work on RAN1, we do not think note 2 is necessary  Mod: Let’s check whether it is acceptable to other companies |
| MediaTek | We have same question as Lenovo, what is **data quality** in the second bullet examples? |
| Futurewei | Support |
| ZTE | Fine with the updated proposal. Per our understanding for data quality, the beam quality (measured RSRP/SINR, etc.) of the data sample may be very weak and impacted most by the noise in some particular cases such as beam failure or blockage. Thus, such non-valid measured data is useless for the model monitoring or fine-tuning and shall be filter out. |
| CMCC | Ok. |

## UE-side AI model training at UE side

In previous RAN1 meeting(s), the agreement(s)/conclusion(s) were made as below:

|  |
| --- |
| **RAN1#111**  Agreement  Regarding the data collection for AI/ML model training at UE side, study the potential specification impact considering the following additional aspects.   * Whether and how to initiate data collection * Configurations, e.g., configuration related to set A and/or Set B, information on association/mapping of Set A and Set B * Assistance information from Network to UE (If supported) * Other aspect(s) is not precluded |

|  |  |
| --- | --- |
| Huawei[2] | *Proposal 8: For the model training/monitoring/inference of the UE-side AI/ML model under BM-Case1 and BM-Case2, for how to indicate the association of beams within Set A and beams within Set B:*   * *Study the indication methods, e.g., indicating the CSI report/resource set ID, time offset, etc.* * *Study the issue when Set A has not been swept in the local cell.*   *Proposal 9: For the model training/monitoring/inference of the UE-side AI/ML model under BM-Case1 and BM-Case2, for how to indicate the mapping of beams within Set A and beams within Set B:*   * *Study the indication methods, e.g., in forms of the set of IDs, bitmap, etc.* * *Study whether/how to indicate such mapping when Set B is a set of wide beams different from Set A.* |
| Vivo[5] | *Proposal 21: Regarding the data collection for AI/ML model training at UE side, study potential specification impact on resource configuration:*  *• Specific beam pair resource configuration for Set A*  *• Enhanced P3+P2 resource configuration that Rx beam assumption of P2 resource measurement is the best Rx beam searched from P3 procedure for performance improvement*  *Proposal 22: Regarding the data collection for AI/ML model training at UE side, study potential specification impact on assistance information:*  *• Proprietary processed Tx beam information as assistance information from NW to UE*  *Proposal 23: Regarding the data collection for AI/ML model training at UE side, study the potential specification impact on request signaling:*  *• Resource request signaling for data collection from UE to NW*  *- Beam pair resources request*  *- P3 and/or P2 beam sweeping resources request*  *• Minimum resource number request for data collection from UE to NW*  *- Minimum number of beams requested for model training w or w/o resource request signaling*  *- Minimum number of repetitions requested for model training w or w/o resource request signaling* |
| OPPO[6] | *Observation 4: For the simplest case of BM-Case1 and BM-Case2 with UE-side AI/ML model trained at UE, data set including model inputs and labels for training can be collected by UE via legacy approach.* |
| Nokia[8] | *Proposal 30. For UE-sided BM-case1 and UE-sided BM-case2, for functionalities supported towards the UE, RAN1 shall study the required CSI-RS measurement enhancements for data collection at the UE side.*  *• Allowing the measurements of Full or partial Set A (associated with a functionality) beam measurements with a longer periodicity than the Set B measurements can be considered.* |
| CATT[9] | *Proposal 2: For DL beam pair prediction with a UE-side model, study the following aspects:*   * *For model training at NW side, study how to report relative Rx beam information when preserving sensitive proprietary information;* * *For model training at UE side, study how to send/report relative Tx beam information when preserving sensitive proprietary information.*   *Proposal 7: Regarding the training data collection for AI/ML model training at UE side, at least study the specification impact on the following aspects:*   * *Indication/request from UE to gNB for RS transmission to be aligned with Rx beam sweeping assumption;* * *Indication of the preferred size of training dataset ;* * *Indication of the minimum periodicity of the RS transmission.* |
| Intel[10] | *Proposal 2: For data collection and UE-side model performance monitoring with AI/ML model at UE side, support UE triggered reference signal transmission from the gNB to enable the UE to perform L1 measurements at least on Set B for both BM-Case 1 and 2.* |
| Ericsson[14] | *Proposal 5 The UE can initiate data collection based on the received configuration/beam ID* |
| Xiaomi[16] | *Proposal 21: For data collection for AI/ML model training at UE side, support UE request for gNB’s configuration with indicating the data size and the preferred relationship between set B and set A.*  *Proposal 22: For data collection for AI/ML model training at UE side, support gNB indicating the relationship between set B and set A.* |
| Samsung[19] | *Proposal 4. For BM-Case1 with a UE-side AI/ML model, for data collection, support the configuration of spatial domain information of Set A and/or Set B, where identifiers can be used for representing Set A beams.*   * *the spatial domain information of Set A and/or Set B should not disclose network implementation*   *Proposal 5. For BM-Case1 with a UE-side AI/ML model, for data collection, support to use the framework of RS configuration and/or CSI report for initiate data collection for UE-side AI/ML model.*  *Proposal 6. For BM-Case1 with a UE-side AI/ML model, for data collection, study the following enhancement:*   * *UE reports it’s the preference on data collection, e.g., preferred RS transmission for measurement, preferred time domain pattern of the RS transmission* |
| CIACT[20] | *Proposal 2(Proposal 3.3 in FL summary): Regarding the training data collection for UE-side AI/ML model trained at UE side, study the potential specification impact of UE reporting of information from the following aspect*  *• Supported/preferred configurations of Resources (e.g., Set A and/or Set B, RS resources)*  *• the number of the needed data samples*  *• Other aspect(s) is not precluded* |
| CMCC[22] | *Proposal 4: Regarding the training data collection for UE-side AI/ML model trained at UE side, study the potential specification impact of UE reporting of information from the following aspect:*   * *Supported/preferred configurations of Resources (e.g., number of beams in Set A and/or Set B, number of RS resources)* * *the number of data samples* * *Other aspect(s) is not precluded*   *Proposal 5: For DL Tx beam prediction with AI/ML model training at UE side, the Rx beam assumption should be aligned between the network and UE.* |
| Lenovo[26] | *Proposal 6: Study data collection procedure to support both UE-side and NW-side AI/ML model training and model update*  * For UE-side model training, study procedure to support UE triggered data collection for model update*  * For NW-side model training, support to report larger number of beams in one beam report.*  *○ FFS: Beam report format, e.g., beam report via MAC CE or RRC* |
|  |  |

###### Proposal 2.3.1(Closed)

Proposal 3.3 of the last meeting were discussed and supported by most companies. According to the tdocs submitted to this meeting, the proposals of most companies are also aligned with this proposal. Thus, we can continue to discuss this proposal. Some minor modifications are made according the discussion of last meeting and the tdocs submitted to this meeting and the new version is provided as Proposal 2.3.1:

* “Regarding the training data collection for xxx” -> “Regarding the training data collection at UE side for”. The main intension is to avoid the potential misunderstanding of dataset delivery from NW.
* “UE reporting of information from the following aspect” -> “UE reporting to network from the following aspect”. The main intension is to emphasize the reporting is from UE to NW

The related proposals in tdocs are as below:

* Vivo: Proposal 23
* CATT: Proposal 7
* Xiaomi: Proposal 21
* Samsung: Proposal 6
* CIACT: Proposal 2
* CMCC: Proposal 4
* Lenovo: Proposal 6

***Proposal 2.3.1: Regarding the ~~training~~ data collection at UE side for UE-side AI/ML model ~~trained at UE side~~, study the potential specification impact of UE reporting to network from the following aspect***

* ***Supported/preferred configurations of DL RS transmission ~~Resources~~ ~~(e.g., Set A and/or Set B, RS resources)~~***
* ***~~The number of the needed data samples~~***
* ***Other aspect(s) is not precluded***

|  |  |
| --- | --- |
| Company | Comments |
| Nokia/NSB | Similar to earlier comments, it does not make sense to collect training data by wasting over-the-air interface resources. The UE can always collect the data based on legacy measurements.  We are fine with the following,  ***Proposal 2.3.1: Regarding the data collection at the UE side, study the potential specification impact of UE reporting to the network from the following aspect***   * ***Supported/preferred configurations of DL RS transmission (e.g., preferred RSs for Set B/Set A, other related conditions)***   Mod: updated. The details can be further discussed. Thus, all examples are removed as more and more examples are suggested by different companies |
| Google | OK with Nokia’s revision. Besides, we think firstly we need to define a framework between the NW and UE to maintain the same understanding on the preferred configurations. Similar to CSI report, the beam management can also be based on a beam codebook.  Mod: This is some kind of the next-level details. Can we discuss it later? |
| New H3C | We suggest deleting bullet 2, as it can be covered by bullet 1, the resource valid period can also be a part of resource configuration.  Mod: Updated |
| LG | OK with Nokia’s version |
| vivo | We are okay with the current formulation of the proposal. |
| Hw/HiSi | We can remove “training” from the main bullet and the support the proposal otherwise:  But when the proposal is not only for training anymore, but also for monitoring or fine-tuning, it might be need to consider further aspects, like a varying number of Top-K output from the inference.  Regarding the comment from Nokia, a compromise for convergence could be following  ***Proposal 2.3.1: Regarding the data collection at the UE side, study the potential specification impact of UE reporting to the network from the following aspect***  ***Supported/preferred configurations of DL RS transmission (e.g., preferred RSs for Set B/Set A, other related conditions such as RS resources, number of needed data samples, FFS: variable Top-K reports)***  Mod: The details can be further discussed. Thus, all examples are removed as more and more examples are suggested by different companies |
| CAICT | We are fine with current version and remove “training” in the main bullet.  Mod: updated. |
| CATT | The minimum periodicity of the RS transmission should also be considered.  The minimum periodicity of RS is related to the UE speed. For example, if the minimum RS periodicity for low-speed Ues is too small, the UE may get the same L1-RSRP results from several subsequent measurements. We suggest the following update:  ***Proposal 2.3.1: Regarding the training data collection at UE side for UE-side AI/ML model trained at UE side, study the potential specification impact of UE reporting to network from the following aspect***   * ***Supported/preferred configurations of Resources (e.g., Set A and/or Set B, RS resources)*** * ***The number of the needed data samples*** * ***The minimum periodicity of the RS transmission*** * ***Other aspect(s) is not precluded***   Mod: it seems included in the 1st bullet as the periodicity is one aspect of the configuration |
| CMCC | Fine with the proposal and agree to remove “training” in the main bullet.  Mod: updated. |
| Ericsson | Support the update from Nokia. Note that the number of needed data samples is very scenario dependent and thus challenging to quantify. |
| Xiaomi | We are fine with the FL’s proposal |
| Samsung | Fine with the revision from Nokia. |
| Fujitsu | Support Nokia’s update. |
| NVIDIA | Support the proposal in principle. |
| Qualcomm | OK with the proposal, given the following suggested updates:  ***Updated Proposal 2.3.1: Regarding the ~~training~~ data collection at UE side for UE-side AI/ML model trained at UE side, study the potential specification impact of UE reporting to network from the following aspect***   * ***Supported/preferred configurations of DL RS transmission (e.g., preferred RSs for Set B/Set A, preferred RS periodicity for Set B/Set A, other related aspects)***   Mod: The details can be further discussed. Thus, all examples are removed as more and more examples are suggested by different companies |
| Lenovo | Suggest taking the second bullet as a sub-bullet of the first bullet.  ***Proposal 2.3.1: Regarding the training data collection at UE side for UE-side AI/ML model trained at UE side, study the potential specification impact of UE reporting to network from the following aspect***   * ***Supported/preferred configurations of Resources***   ***E.g., Set A and/or Set B, resources, the number of the needed data samples***  Mod: The details can be further discussed. Thus, all examples are removed as more and more examples are suggested by different companies |
| MediaTek | We are OK with this proposal. The number of the needed data samples can be reported by UE to NW if UE needs to collect more data for retraining the model.  Mod: As some companies comment, it may be reported implicitly via the 1st bullet. |
| Mod: | The proposal is updated based on inputs. |
| ZTE | Fine with Nokia’s update. |
| Ericsson | Support update |
| Mod | The discussion is closed as a corresponding agreement was made in GTW session. |

###### Proposal 2.3.2(Closed)

In previous meetings, it was agreed to study “Whether and how to initiate data collection”. There are different schemes proposed by companies for which entity can initiate the data collection. In general, these schemes can be categorized into two types:

* Initialed/triggered by NW
* Initialed/triggered by UE

In order to facilitate the related discussion, Proposal 2.3.2 is suggested

The related proposals in tdocs are as below:

* Intel: Proposal 2
* Ericsson: Proposal 5
* Samsung: Proposal 5
* Lenovo: Proposal 6

***~~Proposal 2.3.2: Regarding the training data collection at UE side for UE-side AI/ML model trained at UE side, study the potential RAN1 specification impact (if any) to initial/trigger data collection by considering the following options as a starting point (with potential down-selection)~~***

* ***~~Option 1: data collection initialed/triggered by configuration from NW (e.g., RS configuration, configuration ID)~~***
* ***~~Option 2: the corresponding RS transmission request triggered by UE, including potential operations as below~~***
  + ***~~If gNB confirms the request from UE, gNB and performs the corresponding RS transmission~~***

***Proposal 2.3.2: Regarding the data collection at UE side for UE-side AI/ML model, study the potential specification impact (if any) to initiate/trigger data collection from RAN1 point of view by considering the following options as a starting point***

* ***Option 1: data collection initiated/triggered by configuration from NW***
* ***Option 2: request from UE for data collection*** 
  + ***FFS: details, e.g., if gNB confirms ~~the request from UE~~ UE’s request for RS transmission~~,~~ ~~gNB~~ and performs the corresponding RS transmission***

|  |  |
| --- | --- |
| Company | Comments |
| Nokia/NSB | For model training, this can be left to be discussed by RAN2, specially as they consider framework for data collection.  Mod: As several tdocs mention this issue, we can discuss it in RAN1. In order to avoid the overlapping with RAN2, “RAN1” and “(if any)” are added in the main bullet to focus on RAN1 work. |
| Google | Agree with Nokia  Mod: Please see the reply to Nokia |
| New H3C | We are fine with this proposal in general |
| LG | Agree with Nokia  Mod: Please see the reply to Nokia |
| vivo | We are okay with this proposal. |
| HW/HiSi | Support |
| Spreadtrum | Fine with this proposal. |
| HW/HiSi | Support |
| CATT | Support FL proposal. |
| CMCC | Support. |
| Ericsson | Support FL proposal. Also, ok to leave it to RAN2.  Mod: Please see the reply to Nokia |
| Xiaomi | We are fine with the FL’s proposal |
| Samsung | Similar as earlier comment, there is no need to differentiate the purpose of data collection (e.g., for training or fine-tune). Hence, suggest to remove ‘train’ in the main bullet as below.  ***Proposal 2.3.2: Regarding the ~~training~~ data collection at UE side for UE-side AI/ML model ~~trained at UE side~~, study the potential specification impact to initial/trigger data collection by considering the following options as a starting point (with potential down-selection)***  Mod: updated |
| Fujitsu | Fine to this proposal. |
| NVIDIA | Support the proposal in principle. |
| Qualcomm | Agree with Nokia  Mod: Please see the reply to Nokia |
| MediaTek | We are OK with this proposal. |
| ZTE | Support |
| Ericsson | It is unclear what is meant by “corresponding” in option 2.   * ***Option 2: ~~the corresponding~~ RS transmission request triggered by UE, including potential operations as below***   + ***If gNB confirms the request from UE, gNB performs the corresponding RS transmission***   Mod: Updated |
| Mod | The proposal is replaced by the final version of GTW session. As several companies thought the sub-bullet about gNB operations is important, an FFS part is added. Hope this compromised change can be acceptable to all. |
| MediaTek | We are OK with this version. |
| InterDigital | We are fine with the proposal. |
| Futurewei | OK with the updated version. Propose the following wording change.   * ***FFS: details, e.g., if gNB confirms the request from UE, gNB ~~and~~ performs the corresponding RS transmission***   Mod: this FFS part is updated according to QC’s suggestion. |
| Qualcomm | OK with the following update:  ***Proposal 2.3.2: Regarding the data collection at UE side for UE-side AI/ML model, study the potential specification impact (if any) to initiate/trigger data collection from RAN1 point of view by considering the following options as a starting point***   * ***Option 1: data collection initiated/triggered by configuration from NW*** * ***Option 2: request from UE for data collection***   ***FFS: details, e.g., if gNB confirms ~~the request from UE~~ UE’s request for RS transmission, ~~gNB~~ and performs the corresponding RS transmission***  Mod: updated |
| ZTE | Fine with the proposal. There is a typo in the subbullet as Futurewei mentioned. |
| CATT | Fine with Futurewei’s update. |
| Lenovo | One comment on the FFS point for option 2: Does the gNB need to send a conform message to the UE before transmit the corresponding RS?  Mod: It is up to the further discussion. One possible case is that UE sends a request to NW, and if NW wants to transmit the corresponding RS, it can send a configuration to UE for measurement. This configuration is a kind of implicit “confirm message” in some sense. |
| Xiaomi | We are fine with the updated proposal |
| LG | For Option 2, gNB confirmation or response is somehow natural follow-up procedure but it is highly dependent on what signal is used for the UE request. The confirmation/response may be an existing message/signal that having no spec impact. Thus, we prefer not to add the FFS. Details will be studied after agreeing on this proposal even without any FFS. We see no value to add this FFS.  Mod: Share similar view. Please see the reply to Lenovo. However, as several companies insisted this bullet in previous GTW session. The FFS part is a compromise and should be ok even if there is no much value. |
| Ericsson | Share LG’s concern. The FFS provides limited value.  Mod: Please see the reply to LGE |
| CMCC | Support |
| Nokia/NSB | We are not sure about what refer as “gNB confirms”. We suggest to remove the FFS or change the terminology.  ***Proposal 2.3.2: Regarding the data collection at UE side for UE-side AI/ML model, study the potential specification impact (if any) to initiate/trigger data collection from RAN1 point of view by considering the following options as a starting point***   * ***Option 1: data collection initiated/triggered by configuration from NW*** * ***Option 2: request from UE for data collection***    + ***FFS: details, e.g., ~~if~~ ~~gNB confirms~~ ~~the request from UE~~ ~~UE’s request for RS transmission,~~ ~~gNB and~~ gNB may perform ~~the corresponding~~ RS transmission based on a request coming from the UE*** |
| NTT DOCOMO | We believe that the dataset delivery request is not included in Opt2. Hence, prefer to specify the **RS transmission** request in Option2. |
| HW/HiSi | Ok |
| Mod | The discussion is closed as a corresponding agreement was made in GTW session. |

###### Discussion Point (DP) 2.3.3

There are also some proposals in the tdocs to discuss

* configuration of Set A
* indication of the mapping/association of Set A/B
* Rx beam reporting
* RS configuration enhancement for functionality-based LCM
* …

Since there are usually only one or two companies discussing each issue in tdocs, we can wait for more concrete inputs/proposals from other companies for each of these issues.

|  |  |
| --- | --- |
| Company | Comments |
| HW/HiSi | **Indication of the mapping/association of Set A/B**  We think this is important to be discussed in more detail for the UE-side model.  For a UE-side model, as there may be more than one Set A which the model supports, for a specific Set B which the UE would measure as the input of the AI/ML model, the UE has to be indicated the associated Set A for output from the AI/ML model, for example the specific indication can be the ID of Set A which is included in the configuration for Set B.  This has been agreed for training already in RAN1#112 but should be extended to model monitoring as well, we would should suggest update the agreement below also for model monitoring:  Agreement (RAN1#112)  For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact of AI model inference from the following additional aspects on top of previous agreements:   * Indication of the associated Set A from network to UE, e.g., association/mapping of beams within Set A and beams within Set B if applicable * Beam indication from network for UE reception   Also for UE side model, in legacy, the RS resource set is configured for UE measurement and report; for the AI/ML inference, however, it may not be mandatory for a specific UE to have measured the Set A in prior to the inference, as the training data collection involving the measurement of Set A may occur in other cells or by other Ues. E.g., for the UE handover to a new cell or switch to a new model where Set A is unchanged, gNB may not need to configure the UE to redundantly measure the Set A, and the UE could rather use the stored information of Set A used for training the AI/ML model. In light of this, it should be studied how to tackle the case when Set A has not been swept in the local cell. We have for this the following proposal in our paper.  ***Proposal 8: For the model training/monitoring/inference of the UE-side AI/ML model under BM-Case1 and BM-Case2, for how to indicate the association of beams within Set A and beams within Set B:***   * ***Study the indication methods, e.g., indicating the CSI report/resource set ID, time offset, etc.*** * ***Study the issue when Set A has not been swept in the local cell.***   Another point to consider is that there might be multiple Set B associated with one Set A. If Set B is a subset of Set A, and especially when variable beams are used in Set B (if applicable), as there may be more than one Set B pattern associated with one Set A, the specific pattern of Set B may be indicated to the UE. For this we have the following proposal in our paper.  ***Proposal 9: For the model training/monitoring/inference of the UE-side AI/ML model under BM-Case1 and BM-Case2, for how to indicate the mapping of beams within Set A and beams within Set B:***   * ***Study the indication methods, e.g., in forms of the set of IDs, bitmap, etc.***   ***Study whether/how to indicate such mapping when Set B is a set of wide beams different from Set A***. |
| Mod | Companies are invited to share views on HW’s proposals |
| LG | In general, we also think that beam mapping information should be provided from NW to UE for UE-side model operation (training, inference, monitoring) because UE does not have any information about the relation between beam RSs in current specification. Especially some type of information is needed for UE to know about beams in Set A and beams in Set B for SD prediction. |
|  |  |
|  |  |
|  |  |
|  |  |

# Spec impact of AI/ML inference for BM-Case1 & BM-Case2

## General/common aspects

In previous RAN1 meeting(s), the agreement(s)/conclusion(s) were made as below:

|  |
| --- |
| **RAN1#110**  Agreement  In order to facilitate the AI/ML model inference, study the following aspects as a starting point:   * Enhanced or new configurations/UE reporting/UE measurement, e.g., Enhanced or new beam measurement and/or beam reporting * Enhanced or new signaling for measurement configuration/triggering * Signaling of assistance information (if applicable) * Other aspect(s) is not precluded |

The related proposals/observations are copied as below:

|  |  |
| --- | --- |
| ZTE[4] | *Observation 2: The Tx beam prediction can be achieved by initiating a P2 beam sweeping procedure on data collection for model training and model inference.*  *Observation 3: As the UE Rx beam is up to implementation, the P1 beam sweeping procedure is still conceptual and there is no explicit ignalling/configuration for P1 in current specification.*  *Proposal 2: To facilitate the beam pair prediction, P1 should be specified clearly in the specification with potentially enhanced RS resource set configuration and reporting mechanism.*  *Proposal 6: For Alt.3 (i.e., Set A and Set B are the same) in the beam set construction of BM-Case2, it is useful to configure/transmit the RS resource set (i.e., set A/B) only in the measurement window to minimize the RS overhead for beam measurement.*  *Proposal 7: For Alt.3 (i.e., Set A and Set B are the same) in the beam set construction of BM-Case2, study flexible RS resource set and report configuration within the measurement window and prediction window, regardless of NW-side model or UE-side model.*  *Proposal 8: For Alt.2 (i.e., Set B is a subset of Set A) in the beam set construction of BM-Case2, it is useful to configure/transmit the RS resource set (i.e., set B) in both the measurement window and the prediction window to facilitate a sliding window-based beam prediction.*  *Proposal 9: For Alt.2 (i.e., Set B is a subset of Set A) in the beam set construction of BM-Case2, if set B can be changed along different time instances, study enhanced resource configuration and activation method to flexibly activate/deactivate a beam subset among pre-configured patterns in set A beams (pairs).* |
| Sony[12] | *Proposal 4 : For BM-Case2, the time window size of AI/ML model input can be determined by characteristic of time domain channel.* |
| Xiaomi[16] | *Proposal 1: For BM-Case2, support the periodicity of future time instance is same or 1/N of measurement/report instance.* |
| Google[17] | *Proposal 6: For spatial-domain beam prediction, study to predict the “weak” beam to facilitate the MU-MIMO UE pairing.*  *Proposal 14: For AI/ML based BM, the study should be based on both Rel-17 unified TCI framework and Rel-15/Rel-16 BM framework.* |
| LGE[18] | *Proposal #2: It should be assumed that measurements of Set A beams are not available or is available but can be outdated at UE side, and UE is not expected to manage a good Rx beam for each of the Set A beams.* |
| Lenovo[26] | *Proposal 5: Introduce AI/ML processing units concept for high efficiency AI/ML resource management.*  *Proposal 11: Study on how to obtain the assisting information for AI/ML model input for both NW-side and UE-side AI/ML inference.* |
| DOCOMO[29] | *Observation 1: After the beam prediction, additional beam measurements are necessary for the beam determination in the following case*  *・Top K beam prediction is applied*  *・Rx beam refinement is performed after Tx beam prediction*  *Proposal 2: Study the following two patterns for T1 and T2 in temporal beam prediction.*  *・prediction of beam quality between each measurement/reporting*  *・prediction of beam quality instead of measurement/reporting* |
|  |  |
|  |  |

Most issues are only suggested by one company. Moreover, many proposals are quite general and the concrete details seem not clear. Thus, no proposal is suggested here.

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## AL/ML inference at gNB side

In previous RAN1 meeting(s), the agreement(s)/conclusion(s) were made as below:

|  |
| --- |
| **RAN1#110bis-e**  Working Assumption  For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the following L1 beam reporting enhancement for AI/ML model inference   * UE to report the measurement results of more than 4 beams in one reporting instance * Other L1 reporting enhancements can be considered   **RAN1#111**  Agreement  For BM-Case1 and BM-Case2 with a network-side AI/ML model, study potential specification impact on the following L1 reporting enhancement for AI/ML model inference   * UE to report the measurement results of more than 4 beams in one reporting instance * Other L1 reporting enhancements can be considered |

The related proposals/observations for both BM-Case1 and BM-Case2 are copied as below:

|  |  |
| --- | --- |
| FUTUREWei[1] | *Proposal 1: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study potential specification impact of AI model inference from the following aspects.*   * *Indication of the associated Set A from UE to network, e.g**., association/mapping of beams/beam pairs within Set A and beams within Set B if applicable* |
| Huawei[2] | *Observation 2: For NW-side model of DL Tx beam prediction and BM-Case 1/BM-Case 2, the association/mapping between Set B and Set A is transparent to UE.*  *Proposal 10: For NW-side model of DL Tx-Rx beam pair prediction and BM-Case 1/BM-Case 2, the association/mapping between Set B and Set A may need be studied, where the Rx beam IDs for Set B and the Rx beam IDs for Set A may need to be reported to the gNB.*  *Observation 7: For DL Tx beam prediction when the AI/ML model is at the NW-side, there is no need to introduce additional types of information other than the report of CRI/RSRP, etc., which is already supported by legacy releases.*  *Observation 8: For the AI/ML-based DL Tx beam prediction, non-AI/ML options can be implemented to optimize the Rx beam selection*   * *Opt1: Fixed Rx beams is used for inference during P1/P2 and the Rx beam sweeping is performed to determine the Rx beam in P3* * *Opt2: A quasi-optimal DL Rx beam can be identified by sweeping the always-on SSB beams at P1 and used for Tx beam prediction at P2* * *Opt3: Exhaustive Rx beam sweeping is swept over multiple P1/P2 periods each of which predicts the best Tx beam for a specific Rx beam*   *Observation 9: For DL Tx-Rx beam pair prediction, additional types of beam information are needed for both NW-side model and UE-side model:*   * *When the AI/ML model is located at the UE side, as also for Tx beam prediction, the UE needs to acquire additional types of Tx beam information from the gNB side on top of legacy releases, e.g., Set B pattern involving Tx beams, Set B and Set A association involving Tx beams, etc.* * *When the AI/ML model is located at the NW-side, the NW needs to acquire additional types of Rx beam information from the UE on top of legacy releases, e.g., Set B pattern involving Rx beams, Set B and Set A association involving Rx beams, etc.*   *Observation 10: For DL Tx-Rx beam pair prediction, both when the AI/ML model is located at the NW side or at the UE side, the NW may need to be made aware of the Rx beam number/pattern to interpret the reported Tx-Rx beam pair if the UE reports the Tx-Rx beam pair.*  *Proposal 34: For the inference of the AI/ML model at the NW side, study methods to enable the UE to feedback the RSRP values for a subset of all measured beams in Set B to save UE reporting overhead.*  *Proposal 35: For AI/ML model at the NW-side, no strong motivation to introduce finer resolution for UE reported measurement results at least for model inference.* |
| ZTE[4] | *Proposal 3: For the NW-side beam pair prediction, it is desirable to implicitly indicate the Rx beam ID to facilitate data collection at the NW side and avoid UE proprietary information disclosure issue.*  *Proposal 19: In order to facilitate AI/ML operations for BM-Case1 and BM-Case2, study the following additional aspects:*  *• Beam indication of the unmeasured Tx beam from network to UE*  *• Beam indication of the predicted DL beam pair from network to UE*  *• Beam indication of multiple future time instances for BM-Case2*  *Proposal 20: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the following L1 beam reporting enhancements for AI/ML model inference:*  *• Reporting resolution enhancement*  *• Reporting unmeasured beams*  *• Reporting overhead reduction*  *• Reporting assistance information* |
| Vivo[5] | *Proposal 25: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study potential specification impact on resource configuration for AI/ML model inference:*  *• Specific beam pair resource configuration for Set B*  *• Enhanced P3+P2 resource configuration that Rx beam assumption of P2 resource measurement is the best Rx beam searched from P3 procedure for Set B*  *• Resource configuration for Set C*  *Proposal 26: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study potential specification impact on assistance information for AI/ML model inference:*  *• Proprietary processed Rx beam information as assistance information from UE to NW, including measured Rx beam information, expected Rx beam information, and best Rx beam information.*  *Proposal 27: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study potential specification impact on measurement report for AI/ML model inference:*  *• UE measures the beams of Set B/Set C and reports M1 L1-RSRPs optionally with M2 RS indicators*  *- If M1 is equal to the number of beams or beam pairs in Set B (noted as X), corresponding RS indicators may be not needed.*  *- If M1 is smaller than X/2, corresponding M2 RS indicators are needed*  *- If M1 is smaller than X, but larger than X/2, RS indicators are needed for indicating M2 beams or beam pairs in Set B not included in the measurement report.*  *Proposal 28: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study potential specification impact on report overhead reduction for AI/ML model inference:*  *• Reducing unnecessary L1-RSRP report where the omitted L1-RSRPs may be lower than a pre-defined threshold*  *• Pattern-based beam report if beam resource configuration with multiple pre-configured patterns is supported*  *• Study how to further reduce report overhead of time domain beam prediction for measurement results of multiple occasions.*  *Proposal 29: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study potential specification impact on quantization enhancement for RSRP quality improvement for AI/ML model inference:*  *• High-precision L1-RSRP quantization*  *• Multi-resolution L1-RSRP quantization, e.g. high-resolution quantization for a group of best RSRPs and low-resolution quantization for others.*  *Proposal 30: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study potential specification impact on TCI indication for AI/ML model inference:*  *• Enhanced TCI indication based on both Rel-15/16 and Rel-17 unified TCI frameworks* |
| OPPO[6] | *Proposal 10: For BM-Case1 with NW-side model, study whether/how to reduce the reporting overhead of both fixed or variable Set B, e.g. by dropping the part of SSBRIs/CRIs.*  *Proposal 11: For BM-Case2 with NW-side model, study whether/how UE can assemble multiple instances of Set B measurements into one beam reporting instance.*  *Observation 10: For Tx beam prediction with NW-side model, it seems not necessary to enhance the signaling aspect, e.g. combining or associating the Tx beam indication and the DL Rx beam sweeping.*  *Proposal 12: For BM-Case1 and BM-Case2 with NW-side model, study the feasibility of Tx beam indication and/or beam pair indication.*  *Proposal 13: For BM-Case2 with NW-side model, study the feasibility of beam (pair) indication for multiple future time instance(s) in a single beam indication.* |
| Spreadtrum[7] | *Proposal 2: For sub use cases BM-Case2, implicit indication or report of time information should be considered.*  *Observation 1: For beam indication in BM-Case1 and BM-Case2, the Rel15/16/17 TCI framework can be considered as starting point.*   * *If AI/ML inference is at NW side, the Rx beam indication/selection needs to be enhanced.* * *If AI/ML inference is at UE side, no specification impact is identified*   *Proposal 7: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the enhancement for beam reporting to report one DL Tx beam received by multiple Rx beams.*  *Proposal 8: For BM-Case1 and BM-Case2 with a network-side AI/ML model, existing quantitative criteria should be reused.* |
| Nokia[8] | *Proposal 17. For NW-sided BM-Case1, the following potential specification impact can be considered,*  *• For model inference at the NW, enhancements to the CSI reporting such that the UE can be configured to report the measurement results of more than 4 beams and corresponding L1-RSRP in one beam reporting instance*  *• For beam indication/activation towards the UE, enhancements to the CSI reporting to enable beam measurement and reporting of beams corresponding to the Top-K predicted beams.*  *• For performance monitoring at the NW, study whether enhancements to the CSI reporting are needed to enable full/partial Set A beam measurements.*  *Proposal 18. For NW-sided BM-Case2, the following potential specification impact can be considered,*  *• For model inference at the NW, enhancements to the CSI reporting such that the UE can be configured to report the measurement results of more than 4 beams and corresponding L1-RSRP in one beam reporting instance*  *• For model inference at the NW, enhancements to the CSI measurement and reporting such that the UE can be configured to measure DL RS and report the measurement results for a T1 duration of time and deactivate the measurements/reporting for a T2 duration of time.*  *• For beam indication/activation towards the UE, during T2 duration of time, enhancements to the CSI reporting to enable beam measurement and reporting of beams corresponding to the Top-K predicted beams.*  *• For performance monitoring at the NW, study whether enhancements to the CSI reporting are needed to enable full/partial Set A beam measurements.* |
| CATT[9] | *Proposal 1: For DL beam pair prediction with a NW-side model, study how to report/send relative Rx beam information when preserving sensitive proprietary information.*  *Proposal 10: Regarding the data collection for AI/ML model inference at NW side, UE needs to report the measurement results of more than 4 beams optionally with the corresponding RS indicators to gNB, which can use L1 signaling or MAC CE.*  *Proposal 11: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the following options of beam indication mechanism with potential down-selection:*   * *Opt1: reusing legacy TCI indication mechanism (e.g., Rel-15/16 TCI framework and Rel-17 unified TCI framework);* * *Opt2: only indicate spatial Rx parameter to UE.*   *Proposal 12: For BM-Case2 with a network-side AI/ML model, study how to indicate the beam for multiple future time instances.* |
| Intel[10] | *Observation 2: For beam pair prediction at network side using DL measurements, to align reports from multiple Ues, the network may configure an abstract framework related to spherical coverage through explicit or implicit indication and the Ues can measure DL signals on beams corresponding to configured indexes and report the measurements to gNB which can then construct set B.*  *Proposal 7: For AI/ML model training and inference at NW side, support larger than 4 beams to be reported in one beam reporting instance with potential beam reporting over MAC-CE.* |
| IDC[11] | *Observation 15: The current NR specification supporting UE reporting with up to 4 best CRIs/SSBRIs with L1-RSRP or L1-SINR can be very limited for a network-side AI/ML model.*  *Proposal 21: Consider increasing number of CRIs/SSBRIs (e.g., 8 CRIs/SSBRIs).* |
| Ericsson[14] | *Observation 6 For NW-sided AI/ML model, RAN4 could study the possibility for tightening requirements on the L1-RSRP measurement accuracies for beam prediction use cases*  *Proposal 6 For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the following additional aspects (including the necessity) to facilitate AI model inference:*  *• Quantization of L1-RSRP measurement results*  *• For BM-case 2. Report of compressed value(s) based on temporal sequence of L1-RSRP (e.g. temporal variance or polynomial approximation of L1-RSRP/L1-SINR measurements for beams)*  *• Ues to report the L1-RSRP measurement inaccuracy.*  *• Assistance information (e.g. blockage probability)*  *Note: at least the performance and spec impacts should be considered*  *Note: Assistance information should preserve privacy/proprietary information* |
| Fujitsu[15] | *Proposal 3: Considering the performance of AI/ML model and the overhead of beam reporting, study the necessity and scheme on quantization of beam reporting in the data collection, inference and monitoring on AI/ML model.*  *Proposal 6: Regarding the inference of NW-side model, study the potential specification impacts on the UE behavior of beam reporting.*  *Proposal 7: For the DL beam pair prediction with a NW-side model, study the potential specification impacts on the Rx beam information included in report instance.*  *Proposal 8: Regarding the Rx beam information included in report instance for the DL beam pair prediction with a NW-side model, it is suggested to study*   * *Physical beam information (e.g., beam angle)* * *Logical beam information (e.g., beam ID)*   + *FFS: How to map the logical beam into physical beam* |
| Xiaomi[16] | *Proposal 4: For spatial domain beam prediction, consider to report Rx beam information, including Rx beam ID of UE to gNB for gNB side inference.*  *Proposal 5: For beam indication of Tx beam being not measured by UE, gNB can indicate the Rx beam ID instead of Tx beam ID to UE in the case of Tx/ Rx beam pair prediction at gNB side.*  *Proposal 6: For the case of Tx beam or TxRx beam pair inference with specific Rx, support to indicate Rx beam information to UE for obtaining L1-RSRP input to AI/ML model.*  *Proposal 7: Consider a common AI model for UE with different number of Rx beam.*  *Proposal 9: Support to report predicted L1-RSRP in the L1-beam report with an indication to let gNB know which L1-RSRP is a predicted L1-RSRP.* |
| Google[17] | *Proposal 5: Do not support spec impact for L1-RSRP prediction.*  *Proposal 16: For AI/ML in gNB side, study the following potential enhancement to reduce the L1-RSRP measurement and quantization error.*   * *CSI-RS coverage enhancement* * *More advanced receiver to reduce measurement error* * *High-resolution quantization scheme to reduce quantization error* |
| LGE[18] | *Proposal #7: RAN1 should focus on potential enhancement on L1 beam report, and leave the higher-layer based approach for beam measurement collection to RAN2.*  *Proposal #8: For NW-sided models, in addition to beam reporting enhancements, beam indication enhancement should be considered that TCI/QCL RS should be represented based on Set B beams of which UE can measure and maintain its Rx beam.*  *Proposal #9: For NW-sided AI/ML in BM-Case2, consider enhancements on UE reporting and beam indication.* |
| Samsung[19] | *Proposal 2: For BM-Case1 with a network-side AI/ML model, for model inference, the following aspects should be considered to support a single beam report with more than 4 beams in one reporting instance:*   * *CSI report configuration* * *Content of CSI report* * *Payload size reduction*   *Proposal 14: For BM-Case2 with a network-side AI/ML model, for model inference, study the enhancement of L1 report for future predicted beams:*   * *Enhancement on L1 beam report mechanism (e.g., report more than 4 beams in one reporting instance, enhanced granularity of L1-RSRP)* * *Information about the timestamp corresponding the reported beam(s)*   + *FFS: explicit or implicit* |
| CMCC[22] | *Proposal 6: For BM-Case1 with a network-sided AI/ML model, study the following L1 beam reporting enhancement for AI/ML model inference*   * *How to configure a beam pair pattern from NW to UE* * *UE to Nwwhether/how Rx beam related information corresponding to a measured Tx beam reported from* * *the reporting mechanism enhancement (e.g. differential beam reporting)* |
| MediaTek[23] | *Proposal 3: We support proposal 4.2.2 regarding additional aspects to facilitate AI model inference for a NW-side AI/ML model with the following update:*  *For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the following additional aspects (including the necessity) to facilitate AI model inference:*   * *Quantization of L1-RSRP measurement results*   + *~~The default quantization scheme is NR existing quantization for L1-RSRP reporting if no consensus can be achieved on any other quantization scheme(s)~~*   + *How NW indicates UE the quantization method to use* * *Beam indication of multiple future time instances for BM-Case2*   + *Note: BM-Case2 predicting for near/far-future time instances should be separately discussed*   *Note: at least the performance and spec impacts should be considered* |
| Apple[25] | *Proposal 3: RAN1 should prioritize the study on quantization error’s impact to AI/ML model inference performance over measurement error.* |
| Lenovo[26] | *Proposal 1: Consider the following AI/ML model inputs for both UE-side and NW-side AI/ML inference*  * measured L1-RSRPs corresponding to all the beams within the measurement beam set B with a specific Rx beam are taken as model input for Tx beam ID prediction*  * measured L1-RSRPs corresponding to all the beams pairs which are determined by all the beams within measurement beam set B and all the UE’s Rx beam are taken as model input at least for beam pair prediction*  *Proposal 15: Rel-17 CSI reporting framework can be reused for NW-side beam prediction by increasing the number of beams in a beam report.*  *Proposal 16: To Support NW-side AI/ML inference, the gNB can configured one or more CSI reports for beam report for the UE to report the L1-RSRPs of all the beams configured in the CMR associated with the CSI report.* |
| Qualcomm[27] | *Proposal 4*  *For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the following L1 beam reporting enhancement for AI/ML model inference*  *• Report of temporal and/or spatial variance/variations of L1-RSRP/L1-SINR measurements for beams* |
| NEC[28] | *Proposal 5:* *Regarding model inference for BM-Case2 with NW side AI/ML model, study discontinuous P/SP beam report.* |
| DOCOMO[29] | *Proposal 7: In DL Tx beam prediction with NW side model, configuring Rx beam determination policy, e.g., the same Rx beam for Set B beam measurements, can be considered as potential specification impacts.*  *Observation 6: Enhancements on beam selection policy in CSI reports might be potential specification impacts for spatial domain beam estimation.*  *Proposal 8: L1 beam reporting should be enhanced to facilitate the model inference of temporal beam prediction with the overhead reduction.* |

###### Proposal 3.2.1 (Closed)

Based on the tdocs, the preference of each companies seems not changed. Companies’ views in the tdocs are summarized as below

* Alt.1 (DL Tx beam prediction) is prioritized (at least for NW-side model)
  + Huawei, Ericsson, LGE, Nokia, Samsung, MTK, Apple,
* Alt.3 (DL beam pair prediction) is preferred or support both Alt.1 and Alt.3
  + Futurewei, ZTE, Spreadtrum, OPPO, vivo, IDC, CATT, Fujitsu, CIACT, Intel, DCM(?), Huawei(support DL beam pair prediction for both NW-side and UE-side AI/ML model), Nokia (for some cases)

The opponents raised some issues (especially for the feasibility and the disclosing of UE implementation) of Alt.3 for NW-sided beam prediction. Meanwhile, there are 10 or more companies supporting Alt.3. It seems the only way to move forward is to further study Alt.3 including the feasibility. Thus, based on submitted tdocs previous discussions, a revised version of Proposal 4.2.1 of the last meeting is provided for further discussion.

The related proposals in tdocs are as below:

* Futurewei: Proposal 1
* Huawei: Proposal 10, 12, 13
* ZTE: Proposal 1, 3
* Vivo: Proposal 1, 25, 26
* OPPO: Proposal 12, 19
* Spreadtrum: Proposal 4, 7
* Nokia: 20, 21, 24, 25
* CATT: Proposal 1, Proposal 11
* Intel: Proposal 10
* IDC: Proposal 23
* Fujitsu: Proposal 7, 8
* Ericsson: Proposal 1
* Xiaomi: Proposal 4, 5, 6, 7
* LGE: Proposal 13
* Samsung: Proposal 18
* CIACT: Proposal 1
* CMCC: Proposal 6
* Apple: Proposal 1
* MediaTek: Proposal 3
* Lenovo: Proposal 1
* QC: Proposal 3
* DCM: Proposal 1

***Proposal 3.2.1: For DL beam pair prediction of BM-Case1 and BM-Case2 with a network-side AI/ML model, study the feasibility and potential spec impacts (if feasible) from the following aspects as a starting point***

* ***Whether/How to align the common understanding between NW and UE on the mapping between beam pairs and UE’s associated Rx beams***
* ***Association/mapping of beams/beam pairs within Set A and beams within Set B***
* ***Whether/How to indicate a beam pair / Tx beam /Rx beam from NW to UE***
* ***Whether/How Rx beam related information corresponding to a Tx beam reported from UE to NW***
* ***Generalization aspects, e.g., different UE Rx beam shapes/directions, different UE orientation/location***
* ***Potential assistance information***
* ***Note1: The potential down-selection/prioritization (if any) on the types of beam prediction is a separate discussion***
* ***Note2: The performance, overhead and spec impacts should be considered.***
* ***Note3: Potential reporting and assistance information should not disclose proprietary/privacy information***

|  |  |
| --- | --- |
| Company | Comments |
| Nokia/NSB | We have a problem listing so many study points at this stage of the SI. Do not support this proposal.  We suggest to take out most supported component, and listing them for further study/consider. |
| Google | Agree with Nokia. We think we should firstly study how to maintain the same understanding about the beam related information exchange, e.g., beam grid for set A/B and predicted beam information. We proposed we study the beam codebook to maintain the same understanding for the beams, which is similar to the codebook in CSI. If we cannot reach any consensus on this, we are not sure how to move on especially for UE-side models. |
| New H3C | For sub-bullet 1, we think it’s not needed to align the common understanding between NW and UE, NW should training the model and predict the RX beams based on the reported RX beam IDs which is defined by UE. |
| vivo | OK to study these aspects. |
| HW/HiSi | We propose to prioritize beam prediction over beam pair prediction and do not see a need at this stage for discussion related to spec impact on beam pair prediction, regardless the deployment side. Let’s solve beam prediction firstly, and then see what additionally would be needed.  In evaluations, beam pair prediction has not shown better performance than beam prediction and overhead of beam prediction can also be kept very small (while still having very good performance) if e.g. prior information based on SSB beam sweep is used to select a suitable Rx beam for RSRP measurements.  Furthermore, the evaluations that have been conducted so far for beam pair prediction do not consider non-ideal factors such a UE rotation or Rx beam blocking. This should be included in the simulations. We expect that performance compared to beam prediction will decrease further.  Another aspect is the RSRP measurement error that has not been taking into account yet is the RSRP measurement error. For beam prediction, in most realizations, the RSRPs for different Tx beams are obtained with the same Rx beam, whereas for beam-pair, different Rx beams will be used. In 38.133, clause 10.1.20.2.2, the allowed CSI-RS based relative RSRP measurement error in FR2 is +/- 6.5dB in normal conditions. One of the main reasons for the large error in FR2 is that the Tx beam can be measured with different Rx beams, over which the received signals are separately calibrated and compensated. Thus, a significant larger RSRP measurement error (around 6dB) can be expected for beam pair prediction which will deteriorate the quality of the labels in dataset and the inference output.  Also, the output space of beam pair prediction is larger than for beam prediction. Model complexity is expected to be higher.  We think that before any discussion on spec impact for beam pair prediction (regardless deployment side), the following should be evaluated for beam pair prediction:   * UE rotations and Rx beam blocking (when applicable) * RSRP measurement errors * Complexity   Regarding the sub-bullets proposed here for beam pair prediction, several of them are also applicable for a UE side model (for beam or beam pair):   * ***“Whether/How to align the common understanding between NW and UE on the mapping between beam pairs and UE’s associated Rx beams”***   + ***=>*** This is also needed at the UE side for beam and beam pair, * ***Association/mapping of beams/beam pairs within Set A and beams within Set B***   + ***=>*** This is also needed at the UE side in general for beam and beam pair * ***Whether/How to indicate a beam pair / Tx beam /Rx beam from NW to UE***   + ***=>*** This is also needed at the UE side if Top-K beams are inferred, to signal the NW for Top-K beam sweeping * ***Generalization aspects, e.g., different UE Rx beam shapes/directions, different UE orientation/location***   + ***=>*** if generalization over beams shapes is included, a symmetric issue should be included for UE side model, regardless if it is beam or beam pair prediction. A UE side model should generalize over different Tx beam shapes that can be implemented by different gNB vendors   + ***=>*** UE orientation is an issue that is valid for bother UE and NW. Even if this information might be available somehow at the UE, how to bring it to the AI/ML model (delay, accuracy, etc) would be a different story   If the group anyhow wants to move on with beam pair, prior to further evaluation on beam pair, we think both NW and UE side have to considered equally and that parts of the sub-bullets should be deleted as they apply also to the UE side model for DL Tx beam prediction.  ***For DL beam pair prediction of BM-Case1 and BM-Case2 ~~with a network-side AI/ML model~~, study the feasibility and potential spec impacts (if feasible) from the following aspects as a starting point***   * ***~~Whether/How to align the common understanding between NW and UE on the mapping between beam pairs and UE’s associated Rx beams~~*** * ***~~Association/mapping of beams/beam pairs within Set A and beams within Set B~~*** * ***Whether/How to indicate a beam pair / Tx beam /Rx beam from NW to UE*** * ***Whether/How Rx beam related information corresponding to a Tx beam reported from UE to NW*** * ***~~Generalization aspects, e.g., different UE Rx beam shapes/directions, different UE orientation/location~~*** * ***Potential assistance information*** * ***Note1: The potential down-selection/prioritization (if any) on the types of beam prediction is a separate discussion*** * ***Note2: The performance, overhead and spec impacts should be considered.*** * ***Note3: Potential reporting and assistance information should not disclose proprietary/privacy information*** |
| CAICT | We are fine with the proposal. |
| CATT | For beam pair prediction, the feasibility issue of UE-side model also exists.  For example, if the dataset for UE-side model training comprises the data samples from different gNB Tx beams, how the AI/ML model generalize to different gNB Tx beam shapes/directions should be considered. If the dataset for UE-side model training comprises the data samples from different UE Rx beams (e.g., the model is trained at the third-party server and the dataset consists of beams from multiple Ues), how the AI/ML model generalize to different UE Rx beam shapes/directions should be considered.  Hence, a similar proposal for UE-side model can also be added for discussion. |
| CMCC | Support. Also support to further discuss generalization aspects of Tx beams, e.g., different Tx beam shapes/directions, for both beam pair prediction and Tx beam prediction with UE-side model. |
| Ericsson | We tend to agree with Nokia, this proposal trying to capture an extensive number of aspects. Our preference is to first focus on the feasibility of NW side TX/RX beam predictions. |
| Xiaomi | For the 2nd sub-bullet, it is better to add “from NW to UE” at the end of the sentence.  For the 4th sub-bullet, suggest the following update  ***Whether/How Rx beam related information corresponding to a Tx beam/RSRP reported from UE to NW*** |
| SAMSUNG | We share similar concern with Nokia. In fact, the feasibility of beam pair prediction for NW-side model is far from convincing based on the input in 9.2.3.1. Suggest to postpone the discussion of this proposal. |
| NVIDIA | Support the proposal in principle. It however would be easier for the group to converge if the different parts of the proposal are discussed separately. |
| Qualcomm | We do not support this proposal and agree with Nokia’s assessment. Also, based on the provided comments, it looks like majority (if not all) of the infra vendors are not in favor of NW-side beam pair prediction, or beam pair prediction in general. |
| Lenovo | From our perspective, DL beam pair prediction is lower priority compared with DL Tx beam prediction for NW-side AI/ML model. |
| Apple | We prefer to focusing on study Tx beam prediction over beam pair prediction. |
| MediaTek | We agree with HW’s proposal to prioritize beam prediction over beam pair prediction, we also did not see a need for discussing related spec impact on beam pair prediction at this stage. |
| ZTE | Fine with the FL’s proposal. To move forward, we can first list the potential spec impacts for NW-side beam pair prediction, and then further check the feasibility including how to avoid the UE proprietary/privacy information disclosure issue. |
| Futurewei | We support the proposal in general. Although the items to be studied seem many, we don’t need to do a downselection of them; the items with few proposals in the future can be eliminated. |
| InterDigital | We support the proposal. |

###### Proposal 3.2.2

In the previous meetings, three issues were discussed, but no consensus was achieved.

* Finer granularity
* Indication of unmeasured beams
* Multiple-beam indication

**Finer granularity**

According to the submitted tdocs, companies’ views are unchanged.

* Support: ZTE, Google, vivo, Samsung, MTK, Apple
* Not support: Huawei, Ericsson,

**Indication of unmeasured beams**

According to the submitted tdocs, the number of companies suggesting this study is less than before. Most companies think legacy mechanism can be reused for UE to measure these beams.

**Multiple-beam indication**

A number of companies continue to propose this study.

Considering the above information, moderator suggests to focus on the study of the 1st and 3rd issue. “Necessity” is included in the main bullet to emphasize that there may be no enhancement needed.

The related proposals in tdocs are as below:

* ZTE: Proposal 19, 20
* Vivo: Proposal 29
* OPPO: Proposal 13
* CATT: Proposal 12
* Ericsson: Proposal 6
* Google: Proposal 16
* SS: Proposal 14
* MediaTek: Proposal 3
* Apple: Proposal 3

***Proposal 3.2.2: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study feasibility, necessity, benefit(s) and potential specification impact from the following additional aspects for AI model inference:***

* ***Beam indication of multiple future time instances [in one indication] for BM-Case2***
* ***Measurement reporting of multiple past time instances in one reporting instance for BM-Case2***
* ***Overhead reduction for the reporting of L1-RSRP measurement results (e.g. reporting a partial Set B, L1-RSRP quantization, compressed temporal information for BM-Case2, statistics of past measurements for BM-Case2, etc.)***
* ***Beam indication based on unmeasured/outdated source RS for BM-Case1 and BM-Case2***

|  |  |
| --- | --- |
| Company | Comments |
| Nokia/NSB | What to study is not clear in some aspects.  It would be good to list the exact proposal suggested for Quantization given that no big variation is observed in 9.2.3.1 evaluations on that.  Mod: it can be the work for next step.  Beam indication is always for the future (as long as no new indication, it applies for future instances), not sure what is the suggestion in the second bullet. Can be bit specific here.  Mod: If moderate understood correctly, the proponents are suggesting to use one signaling/indication to indicate the TCI states for multiple future time instance. “in one indication” is added to clarify that. The proponents are encouraged to elaborate a bit more |
| Google | We suggest we revise it as follows:  ***Proposal 3.2.2: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the following additional aspects (including the necessity) to facilitate AI model inference:***   * ***High-resolution quantization scheme for L1-RSRP report*** * ***Coverage enhancement for DL RS for L1-RSRP to reach a measurement accuracy at low SINR for Set B beams***   + ***Taking current measurement accuracy defined in RAN4 as the starting point*** * ***TCI activation and indication with dynamic configured action delay(s) for BM-Case2*** * ***Note: Corresponding evaluations (if any) will be discussed in Agenda item 9.2.3.1***   Mod: The original proposal only includes the issues supported by a number of companies. Companies are invited to share views on the new issue. |
| NTT DOCOMO | For BM Case 2, the hierarchical beam measurements are necessary to track the temporal beam quality change. Then, it is beneficial to support beam reporting of multiple time instances from UE to NW in one reporting instance so that NW obtains the beam measurements of multiple time instances at once from a single reporting. So we suggest to update as follows.  ***Proposal 3.2.2: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the following additional aspects (including the necessity) to facilitate AI model inference:***   * ***Quantization of L1-RSRP measurement results*** * ***Beam indication of multiple future time instances for BM-Case2*** * ***Measurement reporting of multiple time instances in one reporting instance for BM-Case2 model input*** * ***Note: Corresponding evaluations (if any) will be discussed in Agenda item 9.2.3.1***   Mod: updated as at least three companies support it |
| LG | Generally fine but beam indication for BM-Case1 should also be studied. Beam indication is for UE to determine its RX/Tx beam. For DL, UE usually use same RX beam for the QCL source and for the target channel. If the QCL source RS measurement is outdated or unmeasured, e.g. one beam from Set A, UE will have ambiguity on deciding its Rx beam. Thus, we propose to add the following bullet.  • ***Beam indication based on unmeasured/outdated source RS for BM-Case1 and BM-Case2***  Mod: As we have discussed this issue several meetings, it seems difficult to achieve any consensus. This is why it is removed in this proposal |
| vivo | We are okay to list the two points for study. We also think it is needed to spell out more details on these two issues, but these further details can be discussed in subsequent rounds/meetings after agreeing on this proposal from FL. |
| HW/HiSi | For the first sub-bullet, we think what we could focus on is the RSRP reporting overhead. A coarser quantization of the L1-RSRP measurement error could be seen as one tool to reduce the RSRP overhead. But other options should then be listed here as well, like not reporting all beams from Set B.  The second sub-bullet is only applicable for BM-Case 2and should be treated as a separate proposal in our view.  ***Proposal 3.2.2: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the following additional aspects (including the necessity) to facilitate AI model inference:***   * ***Overhead reductions ~~Quantization~~ of L1-RSRP measurement results (e.g. reporting a partial Set B, L1-RSRP quantization, etc.)***   A question for clarification on the 2nd sub-bullet: Does it mean multiple TCI states, or does it include multiple sets of Top-K beams where each would be applied for one prediction instance.  Mod: updated, but the examples are not included. |
| Spreadtrum | For the second bullet, if the gNB indicates multiple beams of multiple instances through the multiple legacy indication. There may be no spec impact, so we recommend adding a note about whether to indicate multiple beams in one beam indication  Mod: please see the reply to Nokia |
| CAICT | We are fine with FL’s proposal and HW’s update for sub-bullet one. |
| CATT | Support the proposal. |
| CMCC | Support. |
| Ericsson | Support the updates by Doccomo followed by HW with a slight update below. Also potential assistance information should be captured.   * ***Overhead reductions ~~Quantization~~ of L1-RSRP measurement results (e.g. reporting a partial Set B, L1-RSRP quantization, compressed temporal information for BM-Case-2, etc.)*** * ***Measurement reporting of multiple past time instances in one reporting instance for BM-Case2 model input*** * ***Assistance information (e.g. blockage probability)***   Mod: If moderator remembered correctly, we have agreement to study assistance information. The details (the examples) can be discussed further |
| Xiaomi | We are with HW’s update for the first sub-bullet. |
| Samsung | Agree with the suggestion from Nokia. More detailed proposal can be provided.  Mod: please see the reply to Nokia |
| SONY | This proposal is fine for us. But for the second bullet, if it should be “predicted beam indication”?  Mod: it seems so. |
| Fujitsu | Support the first sub-bullet. For second sub-bullet, it’s better to be separated to another proposal. |
| NVIDIA | Support. |
| Qualcomm | The proposal seems to be very high-level (both bullets). Based on our evaluations in 9.2.3.1 so far, there’s not a strong motivation for the impact of novel quantization scheme in beam prediction accuracy.  With this being said, it would be helpful if both of the bullets are a bit more specific. |
| Lenovo | We are fine with DOCOMO’s version. |
| MediaTek | Based on our observations and some of companies’ 9.2.3.1 Tdoc, quantizing the L1-RSRP with lower number of bits can lead to minor performance drop compared to legacy quantization scheme. To save the reporting overhead, we think it is worth studying the use of lower quantizing granularity than legacy quantization scheme for L1-RSRP measurement report, especially when UE needs to report L1-RSRPs for >4 beams for NW-side model inference.  Also, we think “facilitate” is misleading as we think this proposal is for both UE reporting and beam indication. To be more specific, we suggest the following word changes:  ***Proposal 3.2.2: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the following additional aspects (including the necessity) ~~to facilitate~~ for AI model inference:***   * ***Quantization of L1-RSRP measurement results for UE reporting*** * ***Beam indication of multiple future time instances for BM-Case2***   ***Note: Corresponding evaluations (if any) will be discussed in Agenda item 9.2.3.1***  Mod: Updated |
| Mod | The proposal is updated based on the inputs.  @all companies: Would you like to check whether the proposal from Google is acceptable or not:  • Coverage enhancement for DL RS for L1-RSRP measurement |
| ZTE | We are fine with DCM’s update. For BM-Case2, the measurement results in the observation window can be reported in one reporting instance, which is beneficial for reporting overhead reduction. |
| Ericsson | Support the update. Although we prefer to remove “Quantization of L1-RSRP measurement results for UE reporting”, this should anyway be captured in the reporting overhead reduction bullet.  Mod: For a reporting, the spec should specify which quantization mechanism is used. Even if the existing L1-RSRP quantization is reused, we still need an agreement/conclusion. The current wording “quantization of the measurement results” is quite inclusive and doesn’t imply any specific mechanism.  Coverage enhancement for DL RS for L1-RSRP should first be more evaluated in 9.2.3.1 prior to including it in other aspects. So far, only Ericsson and Doccomo have studied the measurement sensitivity performance impact. |
| NEC | Fine with the updated version. |
| Samsung2 | For the newly added bullet, we have the following suggestion to make it more generic.   * ***Measurement reporting of multiple past time instances in one reporting instance for BM-Case2 ~~model input~~***   Mod: updated |
| Hw/HiSi | We are ok with most of it but have one question/concern on the second sub-bullet.  In case the NW side model infers multiple sets of Top-K beams, it is not clear that if it is covered by the current proposal that different sets of Top-K beams should be indicated to be swept during different prediction instances. Also, it would maybe be too restrictive to include all of the sets in one indication. To also cover Top-K inference for BM Case 2 (which is very useful as shown in evaluations), we suggest to rephrase the second sub-bullet:  “***Top-1/Top-K*** ***Beam indication(s) of multiple future time instances ~~in one indication~~ for BM-Case2”***  Or is the intention of the second sub-bullet something different than our understanding, could this then be explained please? |
| MediaTek | We are OK with this version |
| InterDigital | Fine |
| Futurewei | OK to us. |
| Qualcomm | We did have a prior item that we had proposed, but we agreed not to add more examples for the sake of progress. Since then, more examples have been added, such as   * ***Measurement reporting of multiple past time instances in one reporting instance for BM-Case2 model input*** * ***Overhead reduction for the reporting of L1-RSRP measurement results***   For the first bullet, do we have evaluation results showing the usefulness of the proposal particularly given the increased overhead?  Mod: Detailed evaluation assumption and KPI can be discussed in EVM session. The main bullet contains “including the necessity”. Evaluation result is one of the key aspects the groups need to consider.  This increased overhead is important to discuss because we are talking about *inference*. Our earlier proposal mentioned above was trying to curb this concern by reporting statistics related to past L1-RSRP measurements rather than the measurements themselves.   * Report of temporal and/or spatial variance/variations of L1-RSRP/L1-SINR measurements for beams   Mod: it is included by overhead reduction, which is more inclusive. There are also some other proposals to reduce the overhead. Thus, a general description on overhead is listed as a sub-bullet. |
| ZTE | Support the updated proposal from FL. The coverage enhancement for DL RS for L1-RSRP measurement should be evaluated first in agenda 9.2.3.1. |
| Lenovo | Fine with the updated proposal. |
| Xiaomi | We think the 1st sub-bullet: quantization of L1-RSRP measurement results was covered by the 4th sub-bullet: overhead reduction, so we suggest to remove the 1st sub-bullet. If no new quantization is introduced, the legacy will be the default.  Mod: Please see the summary on “Finer granularity”. Some companies have other proposals. We can make some conclusion on its necessity based on evaluation results/observations. |
| LG | Again, we propose to add the following bullet. This issue is essential part for SD prediction and mentioned by multiple companies’ tdocs. How UE can determine its beam without knowing the beam for the QCL/TCI source RS? This is fundamental question/issue.  • ***Beam indication based on unmeasured/outdated source RS for BM-Case1 and BM-Case2***  Mod: added |
| Nokia/NSB | * the first bullet is not justified yet in 9.2.3.1. We can check that later if there is justification on changing the quantization used for L1-RSRP having significant concerns. * the fourth bullet, overhead reduction, it is not clear what exactly should be studied as NW-sided model needs additional beam measurements. 9.2.3.1 have not done any evaluations on overhead reduction schemes when reporting Set B measurements. * Beam indication of not measured beam is not clear. Do we foresee that UE will be happy with getting a beam indication where that source RS is not measured at all. What makes sense to consider a case that timeline of the measurement may be longer than what we have now in “known” TCI state definitions of RAN4. In any case, that seems to be a RAN4 discussion. * The last bullet, better to discuss this in 9.2.3.1 than saying “will be discussed in 9.2.3.1”. At this stage, better to evaluate first, before just listing some aspects in the potential spec impacts.   ***Proposal 3.2.2: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the following additional aspects (including the necessity) ~~to facilitate~~ for AI model inference:***   * ***~~Quantization of L1-RSRP measurement results for UE reporting~~*** * ***Beam indication of multiple future time instances in one indication for BM-Case2*** * ***Measurement reporting of multiple past time instances in one reporting instance for BM-Case2 ~~model input~~*** * ***~~Overhead reduction for the reporting of L1-RSRP measurement results~~*** * ***Feasibility of Beam indication based on ~~unmeasured/~~outdated source RS for BM-Case1 and BM-Case2*** * ***~~Note: Corresponding evaluations (if any) will be discussed in Agenda item 9.2.3.1~~*** |
| Mod | The proposal is updated   * The first bullet is deleted. There is an agreement for the quantization in Feifei’s session. Thus, there is no much motivation to introduce finer granularity quantization. * The note is removed   @Nokia: The main bullet is modified as “***study feasibility, necessity, benefit(s) and potential specification impact …***”. Thus, the feasibility of the 5th bullet is included  @Proponents of the 2nd bullet (e.g., ZTE, OPPO, CATT, MTK,…), please reply to Huawei’s comment/suggestions |
| ZTE | The intention of the second bullet is to reduce the signaling overhead for beam indication. In BM-Case2, the optimal beams at multiple future time instances can be predicted in advance with the AL/ML model. Thus, more flexible beam indication and subsequent beam switching methods can be studied. For instance, a list of TCI states and corresponding time stamps can be indicated by the same signaling to be applied in multiple future time instances.  For the 5th bullet, we support the intention that the predicted Tx beam may not be explicitly measured by the UE, which makes the associated QCL relation unavailable. A possible solution for this issue is to configure additional aperiodic RS resources for beam measurement over the predicted top-1/K beams. However, we are confusing how to perform beam indication with unmeasured/outdated source RS. Therefore, we suggest the following generic version to replace the 5th bullet.   * ***Beam indication of the unmeasured Tx beam from network to UE, which may or may not have spec impact*** |
| CATT | Regarding HW’s comment of Top-1/Top-K, we think both Top-1/Top-K can be covered by the bullet.  HW’s suggestion of removing the words “in one indication” is fine to us. Moreover, in BM-Case2, the beam indication scheme when SetB=SetA and Set B is subset of Set A can be different. When Set B=Set A, all the predicted beam (pair)s have been measured, but when Set B is subset of Set A, the predicted beams may not be measured before. The detailed indication mechanism of BM-Case2 can be a next-level discussion.  Based on the above, we suggest the following word change: |
| Ericsson | Some comments,  @Nokia: “*overhead reduction, it is not clear what exactly should be studied as NW-sided model needs additional beam measurements. 9.2.3.1 have not done any evaluations on overhead reduction schemes when reporting Set B measurements. “*  We are not sure if Nokia are referring to their own paper, at least Samsung, Huawei and Ericsson have provided results on how one can reduce the overhead in reporting set B (check variable set B option D results).  For clarity, it would be good to divide the bullet list into what is considered as information used as input to the model. We support the intention that the predicted Tx beam may not be explicitly measured by the UE. The wording can be improved, prefer the text from ZTE.  ***Proposal 3.2.2: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study feasibility, necessity, benefit(s) and potential specification impact from the following additional aspects ~~(including the necessity)~~ for AI model inference:***   * ***~~Quantization of L1-RSRP measurement results for UE reporting~~*** * ***Model input information***   + ***Measurement reporting of multiple past time instances in one reporting instance for BM-Case2***   + ***Overhead reduction for the reporting of L1-RSRP measurement results*** * ***Beam indication of multiple future time instances in one indication for BM-Case2*** * ***Beam indication based on unmeasured/outdated source RS for BM-Case1 and BM-Case2*** * ***~~Note: Corresponding evaluations (if any) will be discussed in Agenda item 9.2.3.1~~*** |
| Xiaomi | Support the updated proposal |
| Hw/HiSi | For BM Case 2 as pointed out also by ZTE, the goal is to save overhead. Having longer prediction periods (more prediction instances) and then a slightly larger Top-K saves overall overhead while it still can maintain a good performance. It would therefore be very helpful of multiple sets of Top-K beams could be indicated for the different prediction instances. This does not necessarily need to be done in one instance.  We are suggesting the following update (as in previous round):  ***Proposal 3.2.2: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study feasibility, necessity, benefit(s) and potential specification impact from the following additional aspects ~~(including the necessity)~~ for AI model inference:***   * ***~~Quantization of L1-RSRP measurement results for UE reporting~~*** * ***Top-1/Top-K*** ***Beam indication(s) of multiple future time instances ~~in one indication~~ for BM-Case2*** * ***Measurement reporting of multiple past time instances in one reporting instance for BM-Case2*** * ***Overhead reduction for the reporting of L1-RSRP measurement results*** * ***Beam indication based on unmeasured/outdated source RS for BM-Case1 and BM-Case2*** * ***~~Note: Corresponding evaluations (if any) will be discussed in Agenda item 9.2.3.1~~*** |
| Lenovo | We have similar feeling that indicating beams for multiple future time instances in one indication is too restricted. We are fine with Huawei’s update on the second bullet. |
| Nokia/NSB | On the overhead reduction part, we should be fine to list it as some companies have done the evaluations and showed some potential.  Minor update as below,  ***Proposal 3.2.2: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study feasibility, necessity, benefit(s) and potential specification impact from the following additional aspects ~~(including the necessity)~~ for AI model inference:***   * ***~~Quantization of L1-RSRP measurement results for UE reporting~~*** * ***Beam indication of multiple future time instances in one indication for BM-Case2*** * ***Measurement reporting of multiple past time instances in one reporting instance for BM-Case2*** * ***Overhead reduction for the reporting of L1-RSRP measurement results*** * ***Beam indication based on ~~unmeasured/~~outdated source RS for BM-Case1 and BM-Case2*** * ***~~Note: Corresponding evaluations (if any) will be discussed in Agenda item 9.2.3.1~~*** |
| Qualcomm | Suggest the following update, to be inclusive of our prior proposal:  ***Proposal 3.2.2: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study feasibility, necessity, benefit(s) and potential specification impact from the following additional aspects ~~(including the necessity)~~ for AI model inference:***   * ***~~Quantization of L1-RSRP measurement results for UE reporting~~*** * ***Beam indication of multiple future time instances in one indication for BM-Case2*** * ***Measurement reporting of multiple past time instances, and/or statistics of past measurements, in one reporting instance for BM-Case2*** * ***Overhead reduction for the reporting of L1-RSRP measurement results*** * ***Beam indication based on unmeasured/outdated source RS for BM-Case1 and BM-Case2***   ***~~Note: Corresponding evaluations (if any) will be discussed in Agenda item 9.2.3.1~~*** |
| CMCC | Fine with Huawei’s update “Top-1/Top-K Beam indication(s) of multiple future time instances for BM-Case2” to enable more inclusive beam indication methods. |
| Futurewei | Support in principle. |
| Mod | The proposal is updated   * Some examples are added to the overhead reduction * “in one reporting instance” is put in bracket since some companies think it is to restricted. Let’s hear more views |
| LG | Supportive in general.   * + Second bullet: ok to add ‘in one reporting instance’ for better clarity   + Third bullet: prefer not to add examples (to save our online time ^^)   Mod: As we can see, several companies continue the requests of some examples. It could be a compromise to move forward.   * + Fourth bullet: support current version. Re Nokia’s suggestion to delete ‘unmeasured’, we slightly prefer to keep it because it may be possible that not all Set A beams are measured by UE before beam indication on Set A from NW depending on NW configuration. |
| Mod | @Ericsson, QC: For the examples “***compressed temporal information for BM-Case2, statistics of past measurements for BM-Case2***”, they seem duplicated. Please correct me if I misunderstand something. |
| CATT | Fine with the updated proposal.  For the third bullet, we also prefer to remove the examples, since none of them are precluded. |
| Samsung | We are a little bit confused with HW’s suggestion. The term of “Top-1/Top-K” is usually used in the context of beam reporting. For beam indication, it seems unnecessary. A wording suggestion is as follows:  Beam indication**(s) applicable** to multiple future time instances **~~in one indication~~** for BM-Case2. |
| Ericsson | @Mod: Agree that they are somehow duplicated. Our view is that compressed temporal information is a bit more general. Proposal for update below, not sure of we need the yellow marked text below let’s hear QCs views:  ***Compressed representation of past measurements for BM-Case2 for example via a statistical value,*** |
| Hw/HiSi | The updated proposal from FL:  We still have a concern on the first bullet, can it be clarified if this addresses multiple top-1 and also top-K instances? And also, if it is top-K, it might not be needed to have the indication in the same reporting instance.  ***Updated: Proposal 3.2.2: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study feasibility, necessity, benefit(s) and potential specification impact from the following additional aspects for AI model inference:***   * ***Top-1/K Beam indication of multiple future time instances [in one indication] for BM-Case2*** * ***Measurement reporting of multiple past time instances [in one reporting instance] for BM-Case2*** * ***Overhead reduction for the reporting of L1-RSRP measurement results (e.g. reporting a partial Set B, L1-RSRP quantization, compressed temporal information for BM-Case2, statistics of past measurements for BM-Case2, etc.)***   ***Beam indication based on unmeasured/outdated source RS for BM-Case1 and BM-Case2*** |
| Mod | The proposal is updated to correct a mistake. The bracket should be in the first bullet, rather than the 2nd bullet   * in one indication -> [in one indication] * [in one reporting instance] -> in one reporting instance   @Huawei: Regarding “Top-1/K”, Let’s hear more views. From my perspective, as “necessity” is included in the main bullet, whether/how to differentiate the designs for Top-1 and Top-K belongs to the next-level details. |
| New H3C | We are fine with the updated proposal. As for the “one reporting instance” in bullet 3, from spec aspect, we think it’s ok. As if one report can’t cover all the past time instances, more report can be sent, which has no spec impact. |
| Vivo | 1. For beam indication of multiple future time instance, we support to have it to reduce signaling latency and overhead. But we would like to clarify that it is only for unified TCI, as for Rel-15/16 TCI framework, one indication is just for one scheduling channel/RS. 2. We suggest to remove the examples for overhead reduction, as apparently there are other examples. To list them is not beneficial at current stage.   Hence we suggest the following revision.  ***Proposal 3.2.2: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study feasibility, necessity, benefit(s) and potential specification impact from the following additional aspects for AI model inference:***   * ***Beam indication of multiple future time instances ~~[~~in one indication~~]~~ for BM-Case2 under unified TCI framework*** * ***Measurement reporting of multiple past time instances in one reporting instance for BM-Case2*** * ***Overhead reduction for the reporting of L1-RSRP measurement results ~~(e.g. reporting a partial Set B, L1-RSRP quantization, compressed temporal information for BM-Case2, statistics of past measurements for BM-Case2, etc.)~~*** * ***Beam indication based on unmeasured/outdated source RS for BM-Case1 and BM-Case2***   Mod: In my understanding, the 1st change should be discussed in WI, rather than SI.  For the 2nd change, let’s check other companies’ view. At least, examples are informative. |
| Spreadtrum | Fine with the updated proposal. |
| MediaTek | We think there is no harm to keep the examples for the overhead reduction bullet. |
| Futurewei | Support |
| ZTE | We support to delete the examples mentioned in the 3rd bullet, which causes confusing and may block the progress since more examples can be added by companies.  For the last bullet, we support the intention that the predicted Tx beam may not be explicitly measured by the UE, which makes the associated QCL relation unavailable. However, whether the unmeasured beam is indicated based on unmeasured/outdated source RS or by triggering an additional aperiodic RS resources for beam measurement over the predicted top-1/K beams can be detailed next step. Thus, we suggest the following revision.  ***Proposal 3.2.2: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study feasibility, necessity, benefit(s) and potential specification impact from the following additional aspects for AI model inference:***   * ***Beam indication of multiple future time instances [in one indication] for BM-Case2*** * ***Measurement reporting of multiple past time instances in one reporting instance for BM-Case2*** * ***Overhead reduction for the reporting of L1-RSRP measurement results ~~(e.g. reporting a partial Set B, L1-RSRP quantization, compressed temporal information for BM-Case2, statistics of past measurements for BM-Case2, etc.)~~*** * ***Beam indication ~~based on unmeasured/outdated source RS~~ of the unmeasured Tx beam from network to UE for BM-Case1 and BM-Case2, which may or may not have spec impact*** |
| CMCC | Support. |
| HW/HiSi | We think it would be good to make it clear whether the inference output is Top-K or Top-1. So far both options are considered in evaluations and Top-K has shown to have better performance. Therefore, I suggest the following small update:  ***Proposal 3.2.2: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study feasibility, necessity, benefit(s) and potential specification impact from the following additional aspects for AI model inference:***   * ***Beam indication (e.g. Top-1 or Top-K) of multiple future time instances [in one indication] for BM-Case2*** * ***Measurement reporting of multiple past time instances in one reporting instance for BM-Case2*** * ***Overhead reduction for the reporting of L1-RSRP measurement results (e.g. reporting a partial Set B, L1-RSRP quantization, compressed temporal information for BM-Case2, statistics of past measurements for BM-Case2, etc.)*** * ***Beam indication based on unmeasured/outdated source RS for BM-Case1 and BM-Case2*** |

###### DP 3.2.1

Some tdocs discuss the indication of Set A/B and/or the association/mapping of Set A/B. However, the detailed proposals are quite diverging. Not find a good proposal to capture this so far.

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## AL/ML inference at UE side

In previous RAN1 meeting(s), the agreement(s)/conclusion(s) were made as below:

|  |
| --- |
| **RAN1#110bis-e**  Agreement  For BM-Case1 with a UE-side AI/ML model, study the potential specification impact of L1 signaling to report the following information of AI/ML model inference to NW   * The beam(s) that is based on the output of AI/ML model inference * FFS: Predicted L1-RSRP corresponding to the beam(s) * FFS: other information   Agreement  For BM-Case2 with a UE-side AI/ML model, study the potential specification impact of L1 signaling to report the following information of AI/ML model inference to NW   * The beam(s) of N future time instance(s) that is based on the output of AI/ML model inference   + FFS: value of N * FFS: Predicted L1-RSRP corresponding to the beam(s) * Information about the timestamp corresponding the reported beam(s)   + FFS: explicit or implicit * FFS: other information   **RAN1#112**  Agreement  For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the necessity, feasibility and the potential specification impact (if needed) of the following information reported from UE to network:   * Predicted L1-RSRP(s) corresponding to the DL Tx beam(s) or beam pair(s)   + Whether/how to differentiate predicted L1-RSRP and measured L1-RSRP * Confidence/probability information related to the output of AI/ML model inference (e.g., predicted beams)   + FFS: Definition/content of confidence/probability information * Note: At least the performance and spec impact should be considered   Agreement  For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact of AI model inference from the following additional aspects on top of previous agreements:   * Indication of the associated Set A from network to UE, e.g., association/mapping of beams within Set A and beams within Set B if applicable * Beam indication from network for UE reception * Note: The second bullet may or may not have additional specification impact (e.g., legacy mechanism may be reused). |

The related proposals/observations for BM-Case1 and BM-Case2 are copied as below:

|  |  |
| --- | --- |
| Huawei[2] | *Observation 1: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, indication of the associated Set A can be used to align the interpretation of the AI/ML output beam(s) from Set A between NW and UE, regardless whether the UE has measured the AI/ML output beam(s).*  *Proposal 8: For the model training/monitoring/inference of the UE-side AI/ML model under BM-Case1 and BM-Case2, for how to indicate the association of beams within Set A and beams within Set B:*   * *Study the indication methods, e.g., indicating the CSI report/resource set ID, time offset, etc.* * *Study the issue when Set A has not been swept in the local cell.*   *Proposal 9: For the model training/monitoring/inference of the UE-side AI/ML model under BM-Case1 and BM-Case2, for how to indicate the mapping of beams within Set A and beams within Set B:*   * *Study the indication methods, e.g., in forms of the set of IDs, bitmap, etc.* * *Study whether/how to indicate such mapping when Set B is a set of wide beams different from Set A.*   *Observation 3: For UE-side model of DL Tx-Rx beam pair prediction and BM-Case 1/BM-Case 2, the association between Set B and Set A is the same as for DL Tx beam prediction and no additional spec impact is needed on top of that.*  *Observation 6: For DL Tx beam prediction when the AI/ML mode is at the UE-side, the UE needs to acquire additional types of information from the gNB side on top of legacy releases, e.g., Set B pattern, Set B and Set A association, etc.*  *Observation 9: For DL Tx-Rx beam pair prediction, additional types of beam information are needed for both NW-side model and UE-side model:*   * *When the AI/ML model is located at the UE side, as also for Tx beam prediction, the UE needs to acquire additional types of Tx beam information from the gNB side on top of legacy releases, e.g., Set B pattern involving Tx beams, Set B and Set A association involving Tx beams, etc.* * *When the AI/ML model is located at the NW-side, the NW needs to acquire additional types of Rx beam information from the UE on top of legacy releases, e.g., Set B pattern involving Rx beams, Set B and Set A association involving Rx beams, etc.*   *Observation 10: For DL Tx-Rx beam pair prediction, both when the AI/ML model is located at the NW side or at the UE side, the NW may need to be made aware of the Rx beam number/pattern to interpret the reported Tx-Rx beam pair if the UE reports the Tx-Rx beam pair.*  *Proposal 31: For AI/ML model inference at the UE-side under BM-Case 1 and BM-Case 2, study the potential spec impact of L1 signaling to report the predicted beam IDs of more than 4 beams in one reporting instance.*  *Proposal 32: For AI/ML model inference at the UE-side under BM-Case 1 and BM-Case 2, study to report an adaptive number of beam IDs/RSRPs determined by the UE, i.e. adaptive values for the Top-K reported beams.*  *Proposal 33: For AI/ML model inference at the UE-side, the motivation of introducing the report of confidence/probability of the AI/ML output is not clear and should be postponed until evaluation results are available in 9.2.3.1.* |
| H3C[3] | *Proposal 5: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, associate/map beams within Set A and beams within Set B based on the Tx beam information:*   * *For DL Tx beam prediction, network can inform UE the beam id of set A together with the Tx beam RS indicator, the mapping of Set B Tx beam with Set A Tx beam can use a bitmap method.* * *For beam pair prediction, network can inform UE the beam id of Set A and set B together with the Tx beam RS indicator.* |
| ZTE[4] | *Proposal 4: For the UE-side beam pair prediction, study enhanced RS configurations to assist UE-side data collection.*  *Proposal 21: Support the reporting of both predicted RSRP and confidence information for beam selection at the NW side.*  *Proposal 22: If both the predicted RSRP and measured RSRP to the same beam are available at the UE side, the measured RSRP should be reported due to its higher reliability.*  *Observation 8: NW can differentiate the predicted RSRP and measured RSRP based on whether or not the reported beam is from Set B.*  *Observation 9: Since the predicted RSRP and measured RSRP of different beams may be reported in one reporting instance, the beam with lower measured RSRP may be associated with higher predicted RSRP in case of prediction error.*  *Proposal 23: Study enhanced reporting mechanism to support the reporting of the predicted RSRP or measured RSRP for different beams.*  *Observation 10: Since the data collection of beam pair prediction incorporates Tx beam sweeping and Rx beam sweeping simultaneously, it may take a long time for one round of data collection for model inference.*  *Proposal 24: Study enhanced resource configuration for P1 beam sweeping procedure to facilitate a timely data collection for model inference of UE-side beam pair prediction.*  *Observation 11: The predicted Top-K beam pairs may include two or more Rx beams that is associated with the same Tx beam.*  *Proposal 25: To differentiate the multiple beam pairs specific to the same Tx beam in the UE reporting, the corresponding Rx beam information can be reported, or the Tx beam associated with different Rx beams can be reported repeatedly.*  *Proposal 26: Considering UE has better knowledge on the confidence level of the predicted top-1 or top-K beams, the additional RS resource for the second stage beam sweeping can be requested by UE.*  *Observation 12: Without Rx beam information reporting, the QCL type D relation associated with each resource for top-K beam sweeping may not be available and thus the RS resource overhead for the second stage beam sweeping may be significantly increased.* |
| Vivo[5] | *Proposal 31: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact on resource configuration for AI/ML model inference:*  *• Specific beam pair resource configuration for Set B/Set C*  *• Enhanced P3+P2 resource configuration that Rx beam assumption of P2 resource measurement is the best Rx beam searched from P3 procedure for Set B/Set C*  *Proposal 32: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact on beam resource request for AI/ML model inference:*  *• Renew beam pattern request w or w/o beam pattern suggestion from UE to NW*  *• Minimum resource number request from UE to NW*  *- Minimum number of requested beams*  *- Minimum number of requested repetitions*  *Proposal 33: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact on assistance information for AI/ML model inference:*  *• Proprietary processed Tx beam information as assistance information from NW to UE*  *Proposal 34: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact on beam report for AI/ML model inference:*  *• Predicted L1-RSRP report and study how to report predicted beam indicator*  *• Fallback beam report to indicate invalid measured results for AI/ML based beam prediction*  *• Study how to further reduce report overhead of time domain beam prediction for predicted results of multiple occasions.*  *Proposal 35: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact on TCI indication for AI/ML model inference:*  *• Enhanced TCI indication based on both Rel-15/16 and Rel-17 unified TCI frameworks* |
| OPPO[6] | *Observation 6: For Tx beam prediction (Alt.1), the corresponding Rx beam could either be determined by UE as specific Rx beam or by existing Rx beam sweeping procedure.*  *Proposal 3: For Tx beam prediction (Alt.1) with UE-side model, legacy beam reporting and indication mechanism could be reused.*  *Observation 7: For beam pair prediction (Alt.3), legacy beam reporting and indication may cause Rx beam confusion (e.g. indicated Tx beam corresponding to different Rx beams).*  *Proposal 4: For beam pair prediction (Alt.3) with UE-side model, study enhanced beam reporting and indication mechanism.*  *• Beam pair reporting: Top-K predicted beam pair info. Along with predicted L1-RSRP(s)*  *• Beam pair indication: both Tx beam and Rx beam info.*  *Observation 8: The reported predicted L1-RSRP and measured L1-RSRP can be differentiated via separate NW configuration.*  *Proposal 5: Predicted L1-RSRP by UE-side model should be reported to NW along with predicted Top-K Tx beam(s) or beam pair(s).*  *Proposal 6: Confidence/probability of UE-side model output could be quantized and reported to NW.*  *Observation 9: From signaling aspects, it seems flexible to configure both Set A and Set B via higher layer signaling.*  *Proposal 7: For BM-Case2 with UE-side model, UE reports the predicted beam (pair) for N future time instance(s) by single reporting instance.*  *Proposal 8: For BM-Case2 with UE-side model, the timestamp of N future time instance(s) should be implicitly reported to NW.*  *Proposal 8: For BM-Case2 with UE-side model, the timestamp of N future time instance(s) should be implicitly reported to NW.*  *Proposal 9: For BM-Case2, NW indicates multiple beam indications for future N time instances.* |
| Spreadtrum[7] | *Proposal 2: For sub use cases BM-Case2, implicit indication or report of time information should be considered.*  *Observation 1: For beam indication in BM-Case1 and BM-Case2, the Rel15/16/17 TCI framework can be considered as starting point.*   * *If AI/ML inference is at NW side, the Rx beam indication/selection needs to be enhanced.* * *If AI/ML inference is at UE side, no specification impact is identified*   *Proposal 9: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the enhancement for beam report without RSRP.* |
| Nokia[8] | *Proposal 22. For UE-sided BM-Case1 with DL Tx-Rx beam pair prediction, study methods to reduce the necessary measurement space for DL TX-RX beam pair prediction at the UE side.*  *• Study the enhancements related to the applicable conditions where UE indicates a number of preferred TX beams along with a number of “P3” repetitions that are needed for each preferred TX beam, in which case the UE can acquire L1-RSRP measurements of the indicated combinations as inputs to the UE-sided AI/ML model.*  *Proposal 27. For UE-sided BM-Case1 a with a UE-side AI/ML model, study the potential specification impact of L1 signaling to report Predicted L1-RSRP to the NW,*  *• RAN1 may further investigate additional applicable conditions for L1-RSRP reporting.*  *Observation 1. To distinguish predicted L1-RSRP from measured L1-RSRP when the UE-sided model is employed,*  *• If a reported beam belongs to Set B, NW knows it is a measured L1-RSRP, otherwise, NW knows it is a predicted L1-RSRP.*  *Proposal 28. RAN1 to consider reporting of confidence/probability information related to the output of AI/ML model inference (e.g., predicted beams).* |
| CATT[9] | *Proposal 10: Regarding the data collection for AI/ML model inference at NW side, UE needs to report the measurement results of more than 4 beams optionally with the corresponding RS indicators to gNB, which can use L1 signaling or MAC CE.*  *Proposal 11: For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the following options of beam indication mechanism with potential down-selection:*   * *Opt1: reusing legacy TCI indication mechanism (e.g., Rel-15/16 TCI framework and Rel-17 unified TCI framework);* * *Opt2: only indicate spatial Rx parameter to UE.*   *Proposal 12: For BM-Case2 with a network-side AI/ML model, study how to indicate the beam for multiple future time instances.* |
| Intel[10] | *Observation 1: For beam pair prediction at UE side using DL measurements, if gNB provides indexes ordered in terms of angular coverage of beams, there may be no need to divulge proprietary information and UE can use the beam indexes to construct set B for input to ML model.*  *Proposal 4: If the predicted RSRP is directly obtained from the output of the AI/ML model, such metrics should not be reported back to the network or should be reported with an indication that the report is a prediction from UE side model and not an actual measurement.*  *Proposal 5: If the predicted RSRP is defined as the actual measured RSRP on the best beam predicted by the model, such metric can be reported back to the network assuming that the overhead due to the additional measurement, if needed, is accounted for in the KPI calculation.* |
| IDC[11] | *Observation 12: For BM-Case 1 with a UE-side AI/ML model, the current beam reporting with CRIs/SSBRIs and corresponding L1-RSRP values is enough to indicate the best beam(s) from AI/ML model inference.*  *Observation 13: For BM-Case 2 with a UE-side AI/ML model, information about the time stamp for the reported CRIs/SSBRIs can be further considered.*  *Proposal 18: For a UE-side AI/ML model, consider information about the time stamp for potential specification impact.*  *Observation 14: Reporting confidence/probability information per predicted beam can cause unnecessary reporting overhead.*  *Proposal 19: Consider reporting confidence/probability information related to the output of AI/ML model per an output instance.*  *Proposal 20: Consider using legacy procedures to indicate the mapping between Set A and Set B to the UE.* |
| Panasonic[13] | *Proposal 1: CSI reporting framework can be considered as starting point for UE to report beam prediction to NW in case of UE-side inference.*  *Proposal 2: Prediction related metrics can be introduced in the CSI report configuration as the report quantities. FFS the following prediction related metrics:*   * *Predicted beam ID (or RS ID, or TCI State ID)* * *Predicted beam quality, such as predicted L1-RSRP, L1-SINR* * *Predicted beam application time (when to start/stop applying the predicted beam)* * *Confidence/probability information*   *Proposal 3: RAN1 to discuss mechanism for NW to distinguish between prediction and measurement results.*  *Proposal 4: RAN1 to discuss beam grid approach to indicate the mapping relationship among beams (for measurement and prediction) to the UE.* |
| Ericsson[14] | *Proposal 2 Conclude that the specification impact for DL beam pair prediction at UE sided model inference is same as for TX DL beam prediction*  *Observation 7 Depending on the AI/ML model, confidence can be estimated:*  *a. DuriEng inference, confidence is dependent on the input.*  *b. During training, confidence is constant for all inputs during inference.*  *Proposal 7 For the input-dependent confidence reporting during UE-sided AI/ML inference, study feasibility and specification impact for the following alternatives:*  *a. Probability/likeliness of strongest beam for each Top-K beam*  *b. Confidence interval (e.g. 95th percentile) for L1-RSRP prediction for a predicted beam*  *Proposal 8 For constant confidence reporting for all input (based on the training step) for UE-sided AI/ML inference, study the feasibility and specification impact for the following alternatives:*   1. *Strongest beam prediction (log-loss, accuracy,*   *L1 RSRP error in e.g. 50th and 95th percentile to the genie aided beam)*  *b. L1-RSRP prediction (e.g. L1-RSRP error in e.g. 50th and 95th percentile for the predicted Top-1/K beam)*  *Proposal 9 For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact of AI model inference from the following additional aspects on top of previous agreements*  *a. Enhanced CSI resource/report configuration, e.g. how to adapt the TCI switch time offsets or configure several TCIs in one configuration.*  *b. Beam indication for a UE to switch to a predicted beam with unknown TCI state.*  *Proposal 10 In order to facilitate AI/ML operations for BM-Case1 and BM-Case2 with UE-side AI model, study the necessity and the potential specification (if needed) of the following aspect on data collection for training, inference and/or monitoring:*  *• Scenario identification from NW to UE (e.g. antenna/beam configuration IDs)*  *Observation 8 The number of unique beam IDs should be restricted to minimize the signalling overhead of beam IDs* |
| Fujitsu[15] | *Proposal 9: For DL beam (pair) prediction with a UE-side model, study the potential specification impacts of model inference on*   * *The request to NW about the required Tx beams of Set B* |
| Xiaomi[16] | *Proposal 11: Consider one absolute L1-RSRP for each time instance or one absolute L1-RSRP for all time instance in one beam report including beam reports of more than one time instance for BM-case 2.*  *Proposal 12: Consider UE to report the number/ periodicity of the time instance in beam report for BM-case 2.* |
| Google[17] | *Proposal 17: For AI/ML in UE side, study the potential enhancement to maintain the same understanding between the gNB and UE with regard to the reported beam information based on a beam-codebook similar to CSI feedback based on a codebook*   * *The UE can report a beam matrix indicator (BMI) based on the beam-codebook* |
| LGE[18] | *Proposal #3: Consider UE assistance/reporting for determining Set A, e.g. UE to report preferred Set A among candidate beams of Set A.*  *Proposal #4: Support predicted L1-RSRP report together with beam(s). For BM-Case2, information on time-variation of L1-RSRP can also be included in the report for helping intra-/extra-polation at NW side.*  *Proposal #5: For predicted L1-RSRP report, confidence/probability information may be helpful for NW to decide whether/how to use the reported L1-RSRP. Further study whether the information is per model/functionality, per report or per report parameter.*  *Proposal #6: For BM-Case2 with UE-sided models, following beam reporting enhancements can be considered*   * *Report of beam(s) for each future time instance or beam(s) for a time duration, i.e. from the first time instance to the last time instance.* * *Report of beam(s) for current time instance for fallback operation* * *Report of timestamps by UE or NW to indicate timestamps* |
| Samsung[19] | *Proposal 7. For BM-Case1 with a UE-side AI/ML model, for model inference, support the configuration of spatial domain association of Set A and/or Set B, where identifiers are needed for representing Set A beams.*   * *the spatial domain information of Set A and/or Set B should not disclose network implementation*   *Proposal 8: For BM-Case1 with a UE-side AI/ML model, for model inference, further study the specification impacts on the following aspects:*   * *UE to report the predicted beam using the identifiers for Set A beams*   *Proposal 9: For BM-Case1 with a UE-side AI/ML model, for model inference, further study the feasibility and specification impacts on the following aspects:*   * *Predictive beam indication*   *Proposal 10: For BM-Case1 with a UE-side AI/ML model, for model inference, further study the feasibility to predicted L1-RSRP and confidence or probably information corresponding to a predicted beam.*  *Proposal 15: For BM-Case2 with a UE-side AI/ML model, for model inference, study the enhancement of L1 report for future predicted beams:*   * *For the beam(s) of N future time instance(s), N = 1 is baseline* * *Implicit timestamp corresponding the reported beam(s) is baseline*   *Proposal 16: For BM-Case2 with a UE-side AI/ML model, for model inference, further study the feasibility to predicted L1-RSRP corresponding to a predicted beam.* |
| ETRI[21] | *Proposal 2: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, consider the following as the baseline for the information reported from UE to NW:*  *- The probability information associated with the AI/ML model output can be defined as the probability of each beam being identified as the optimal beam.*  *- In terms of reporting information, there may not be a need for individual content for confidence in addition to probability.*  *Proposal 3. For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact of the approach that distinguishes the relationship between Set A and Set B using AI/ML model’s ID.* |
| CMCC[22] | *Proposal 7: For BM-Case1 with a UE-sided AI/ML model, study the following L1 beam reporting enhancement for AI/ML model inference*   * *How to configure a beam pair pattern from NW to UE* * *whether to support UE to report the measurement results of more than 4 beams (pairs) in one reporting instance* * *whether Rx beam related information corresponding to predicted top K beam pairs reported from UE to NW* * *If UE does not report Rx beam related information, additional spec impact compared to DL beam prediction*   *Proposal 8: For BM-Case1 with a UE-side AI/ML model, whether the predicted L1-RSRP is reported can be configured by the gNB, whether/how to differentiate measured L1-RSRP and predicted L1-RSRP needs further discussion.* |
| Lenovo[26] | *Proposal 10: Rel-17 CSI reporting framework can be reused for UE-side beam prediction by configuring measurement beam Set B as the channel measurement resource, but the reported beam is selected from another prediction beam Set A.*  *Proposal 13: Study the mechanism for beam report associated with AI/ML inference when there is no available AI/ML model for AI/inference.*  *Proposal 14: For a beam report associated with AI/ML inference, the UE indicate that the reported beams are predicted beams or measured beams in the beam report.* |
| Qualcomm[27] | *Proposal 2*  *For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the following aspects related to association/mapping of beams within Set A and beams within Set B which is indicated from NW to UE:*    *• QCL relation between beams within Set A and beams within Set B*  *• Beams within Set B are superposition and/or linear combination of beams within Set A (e.g., for wide-to-narrow beam prediction)*  *• Relative beam pointing angles of beams within Set A and beams within Set B*  *• FFS: other options*  *Proposal 5*  *For BM-Case1 with a UE-side AI/ML model, study the potential specification impact of L1 signalling to report the following information of AI/ML model inference to NW:*  *• Information about NW DL TX beam angles from target prediction beam set (Set A)*  *• UE may predict best beam angles from target prediction beam set (Set A) by measuring measurement beam set being input to AI/ML model (Set B) of DL TX beams*  *• FFS: details of beam angle, e.g., beam boresight direction*  *Proposal 6*  *For BM-Case2 with a UE-side AI/ML model, study the potential specification impact of L1 signalling to report the following information of AI/ML model inference to NW:*  *• Predicted beam blockage/failure* |
| NEC[28] | *Proposal 3: Support selecting Top-N1 DL Tx and/or Rx beams according to some pre-defined rules, e.g., a sum* *probability of being the best beam higher than a threshold, L1-RSRP higher than a threshold.* |
| DOCOMO[29] | *Proposal 10: Study two-stage beam measurements with top-N predicted beams, since it reduces RS measurement overhead and increases the reliability of beam selection compared to top-1 beam prediction.*  *Observation 7: Time instances corresponding to the reported beam(s) does not need to be reported when CSI reference resource is aligned between UE and NW, since it can be implicitly determined based on the measured RS occasion and the time offset.*  *Observation 8: It is beneficial to report the explicit predicted time instances in the reporting, if CSI reference resource is not always aligned between UE and NW.* |

###### Proposal 3.3.1 (Closed)

In the last meeting, Proposal 4.3.1 was suggested in order to preserve the proprietary information and keep a focused scope. Based on the submitted tdocs, many companies propose enhancements for the beam pair prediction. One possible way to move forward is to identify the potential additional spec impact and then decide whether a given enhancement is needed or not. Thus, Proposal 3.3.1 is suggested to encourage companies to provide more concrete inputs.

The related proposals in tdocs are as below:

* Huawei: Observation 6, 9, 10
* ZTE: Proposal 4
* Vivo: Proposal
* OPPO: Proposal 4
* Spreadtrum: Observation 1
* Nokia: Proposal 22
* CATT: Proposal 10
* Intel: Observation 1
* Ericsson: Proposal 2
* Fujitsu: Proposal 9
* CMCC: Proposal 1
* NEC: Proposal 3

***Proposal 3.3.1: For DL beam pair prediction of BM-Case1 and BM-Case2 with a UE-side AI/ML model, RAN1 study (including necessity and feasibility) and, if any, identify the potential additional spec impact(s) compared to DL beam prediction, with the following aspects as a starting point***

* ***…***
* ***Note1: Privacy/proprietary information should be preserved***
* ***Note2: Performance and overhead (e.g., RS overhead, reporting overhead) should be considered***

|  |  |
| --- | --- |
| Company | Comments |
| Mod | The current version of Proposal 3.3.1 is suggested to encourage companies to provide more concrete inputs. Refinement(s) will be made based on the inputs. |
| Nokia/NSB | Compared to DL Tx beam prediction, we foresee that the applicable conditions associated with functionalities to indicate a number of preferred TX beams along with a number of “P3” repetitions that are needed for each preferred TX beam |
| Google | As we commented before, the DL RS coverage could be a problem to meet the measurement accuracy. |
| Vivo | For UE side AI/ML model of beam pair prediction, we think the following aspects can be different from DL beam prediction.   * RS configuration for the measurement * Signaling for association between output and input |
| HW/HiSi | At this point, there is no need to discuss the spec impact specifically for beam pair prediction at UE side since several aspects would need to be evaluated firstly (e.g. generalization to different Tx beam shapes, UE rotations, Rx beam blocking, RSRP measurement error, model complexity).  For more details, please see our comments to proposal 3.2.1 (if the group wants to move on, we prefer a common agreement for UE side and NW side deployment) |
| CATT | Agree with Nokia. |
| CMCC | If DL beam pair prediction is feasible, the beam pair pattern configuration, the indication of Top K beam pair from UE to NW can be studied. |
| Ericsson | Agree with HW |
| Samsung | In our view, the proposal should be formulated in a purpose-specific manner. For example, for the purpose of data collection, we are fine with the suggestion from Nokia. |
| NVIDIA | This appears to be an empty proposal? It’s more like a FL recommendation.  Mod: As explained in the beginning of this table, the current version is used to collect detailed inputs, not a “real” proposal. |
| Qualcomm | Agree with HW |
| Apple | We are not convinced by the performance/benefit/necessity of beam pair prediction. |
| MediaTek | Agree with HW |
| ZTE | Since the data collection of beam pair prediction incorporates Tx beam sweeping and Rx beam sweeping simultaneously, it may take a long time for one round of data collection for model inference. Therefore, we need to study enhanced resource configuration for P1 beam sweeping procedure to facilitate a timely data collection for model inference of UE-side beam pair prediction. Besides, we note that the predicted Top-K beam pairs may include two or more Rx beams that is associated with the same Tx beam.Therefore, to differentiate the multiple beam pairs specific to the same Tx beam in the UE reporting, the corresponding Rx beam information can be reported, or the Tx beam associated with different Rx beams can be reported repeatedly.  Therefore, we support the following two aspects for further study.   * enhanced resource configuration for P1 beam sweeping procedure   how to differentiate the multiple beam pairs specific to the same Tx beam in the UE reporting |
| NEC | Agree with HW |

###### Proposal 3.3.2

According to the tdocs, two issues are mentioned/discussed by several companies.

* How to report and differentiate the predicted L1-RSRP and the measured L1-RSRP
* Reporting more than 4 predicted beams and the associated L1-RSRP (if applicable) in one reporting instance

Thus, Proposal 3.3.2 is provided for further discussion.

The related proposals in tdocs are as below:

* Huawei: Proposal 31
* Lenovo: Proposal 14
* CMCC: Proposal 7, 8
* Panasonic: Proposal 3
* Intel: Proposal 4
* Nokia: Observation 1
* ZTE: Proposal 22
* CATT: Proposal 10

***Proposal 3.3.2: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the necessity, benefit(s), and potential specification impact of AI model inference from the following additional aspects on top of previous agreements:***

* ***For BM-Case1: L1 reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) in one reporting instance***
* ***For BM-Case2: L1 Reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) for each one of N time instance(s) in one reporting instance***
  + ***FFS: values of N (e.g., fixed or variable)***
* ***FFS: How to reduce the overhead***
* ***Note1: The performance gains should be justified by considering UCI payload overhead***
* ***~~Note2: The associated L1-RSRP could be predicted or measured L1-RSRP.~~***

|  |  |
| --- | --- |
| Company | Comments |
| Nokia/NSB | Ok in general. |
| Google | Do not support. No benefit for L1-RSRP prediction according to our simulation.  Mod: We have made an agreement to study the necessity of predicted L1-RSRP reporting. This is a separate issue. “if applicable” is added to emphasize that it doesn’t mean predicted L1-RSRP should be supported |
| New H3C | we are ok. |
| LG | Unclear on the benefit of reporting more than 4 beams for UE-side model  Mod: Some companies think that reporting Top-K (K>4) is beneficial for some use cases. Proponents can elaborate a bit more to convince other companies |
| vivo | OK |
| HW/HiSi2 | Ok in principle, with minor modification to control the extra overhead that would be introduced with this proposal. If the number of predicted beams is increased to more than 4, it will cost more overhead. And in many cases this won’t be needed and depends on the inference output of the UE side model.  E.g., if the output probabilities for 64 beams in Set A are distributed as {60%, 35%, 1%, 1%, 1%, 1%, 1%, 0, 0, …}, the UE can feedback K=2 with ~95% probability of achieving the genie-aided Top-1 in total; on the other hand, if the output probabilities in Set A are distributed as {25%, 20%, 20%, 15%, 15%, 1%, 1%, 1%, 1%, 1%, 0, 0, …}, it is more risky for the UE to only report K=2 beams (which contributes just 45% probability in total in this case) while reporting K=5 would bring a higher accuracy.  Also, BM-Case 2, the number of prediction instances might be flexible. Under more constant condition, observations windows can be further apart in time.  ***Proposal 3.3.2: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact of AI model inference from the following additional aspects on top of previous agreements:***   * ***For BM-Case1: L1 reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) in one reporting instance*** * ***For BM-Case2: L1 Reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) for each one of N time instance(s) in one reporting instance***   + ***FFS: values of N***   + ***Note: The number of reporting instances between two observations might vary*** * ***Note: The number of reported predicted beams may vary between different reporting instances.***   Mod: It seems we can discuss these details next step (or even in WI if there is any). The current version is quite inclusive. |
| Spreadtrum | Agree |
| CAICT | Support. |
| CATT | The beams UE reported can be measured (the beams in Set B) or predicted, we suggest the following update:  ***Proposal 3.3.2: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact of AI model inference from the following additional aspects on top of previous agreements:***   * ***For BM-Case1: L1 reporting of more than 4 ~~predicted~~ beams and the associated L1-RSRP (if applicable) in one reporting instance*** * ***For BM-Case2: L1 Reporting of more than 4 ~~predicted~~ beams and the associated L1-RSRP (if applicable) for each one of N time instance(s) in one reporting instance***   ***FFS: values of N***  Mod: It is moderator’s understanding, no matter the beam is measured or not, the reported beams are based on the output of AI model. In this sense, it is a predicted beam. |
| CMCC | Ok. |
| Ericsson | Support FL proposal. We don’t see how one can report a measured beam in BM-Case2 as suggested by CATT. In the predicted beam report, the beams can also have an associated confidence/probability. We therefore suggest the following;  ***Proposal 3.3.2: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact of AI model inference from the following additional aspects on top of previous agreements:***   * ***For BM-Case1: L1 reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) in one reporting instance*** * ***For BM-Case2: L1 Reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) for each one of N time instance(s) in one reporting instance***   + ***FFS: values of N***   ***FFS: Confidence/probability information related to the predicted beams***  Mod: If moderator remembered correctly, the new part is included in previous agreement. |
| Xiaomi | For UE side model, what is the motivation of reporting more than 4 beams per time instance?  Mod: Please see the reply to LG |
| Samsung | Ok. |
| Sony | we are not sure if our understanding is correct, the “associated L1-RSRP” is predicted L1-RSRP and it’s also the model’s output, is right?  Mod: As we discussed in the last meeting, the reported L1-RSRP may be the predicted or measurement L1-RSRP. |
| Fujitsu | For the inference of UE-side model, it’s not clear why UE is required to report more than 4 beams.  Mod: Please see the reply to LG |
| NVIDIA | Support |
| Qualcomm | First, it is not clear how much benefit we would get by reporting more than 4 predicted beams. Second, even if we actually see gains in beam prediction accuracy performance by reporting more than 4 predicted beams, this comes with the price of increased UCI payload overhead. Unless the trade-off between beam prediction accuracy and the UCI payload overhead are not studied and the benefits are not justified in 9.2.3.1, we think the potential specification impact should not be considered.  Mod: Please see the reply to LG |
| Lenovo | Fine |
| MediaTek | We are fine with this proposal. Current L1 reporting already includes the L1-RSRP for each reported beam, therefore, we think it is reasonable to include the associated L1-RSRP. |
| Mod: | The proponents are encouraged to provide more information to convince LGE, Xiaomi, Fujitsu, QC why K<=4 is not sufficient.  “(including necessity)” is added in the main bullet to address the concerns of LGE/Xiaomi/Fujitsu/QC. |
| ZTE | The reporting of more than 4 predicted beams is not needed for the UE-side model. Instead, since the predicted RSRP and measured RSRP of different beams may be reported in one reporting instance, the beam with lower measured RSRP may be associated with higher predicted RSRP in case of prediction error. Therefore, we need to study enhanced reporting mechanism to support the reporting of the predicted RSRP or measured RSRP for different beams. |
| HW/HiSi | @FL Thank you for the feedback  “Mod: It seems we can discuss these details next step (or even in WI if there is any). The current version is quite inclusive.”  We think especially when more than 4 beams are reported the overhead can be rather larger, if constantly reported. Therefore, to give guidance for the next step, we would like to include an FFS that the number of reports is variable or fixed.  ***Proposal 3.3.2: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact (including necessity) of AI model inference from the following additional aspects on top of previous agreements:***   * ***For BM-Case1: L1 reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) in one reporting instance*** * ***For BM-Case2: L1 Reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) for each one of N time instance(s) in one reporting instance***   + ***FFS: values of N, (e.g. fixed or variable)*** * ***FFS: number of reported predicted beams same/different for different reporting instances.***   Mod: The proposal is updated. |
| Panasonic | We support the proposal in general. To our understanding, the “associated L1-RSRP” could refer to predicted or measured RSRP. And we also agree with ZTE that in some cases it could be useful to report both predicted and measured RSRPs. Therefore, we suggest to add one new bullet:   * The associated L1-RSRP could be predicted and/or measured RSRP. FFS: whether and how to differentiate predicted and measured L1-RSRP.   Mod: According to the legacy mechanism, the measured L1-RSRP can be reported. We have the following agreement in previous agreement. Thus, it seems the new bullet has been covered.   * Predicted L1-RSRP(s) corresponding to the DL Tx beam(s) or beam pair(s)   + Whether/how to differentiate predicted L1-RSRP and measured L1-RSRP |
| NEC | Agree with HW. From our point of view, the best beams to be reported should be determined based on the predicted L1-RSRPs or probabilities of being the best beam. Due to the uncertainty of the predicted L1-RSRP or probability, number of predicted beams to be reported for different reporting instances may be variable.  Mod: The proposal is updated. |
| Mod | The proposal is updated by adding the last three red parts |
| HW/HiSi | We are ok with the direction of the proposal, but think that the Note should be deleted. It is not sure that the overhead will become larger considering, that we have the FFS how to reduce the overhead.  ***Proposal 3.3.2: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact (including necessity) of AI model inference from the following additional aspects on top of previous agreements:***   * ***For BM-Case1: L1 reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) in one reporting instance*** * ***For BM-Case2: L1 Reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) for each one of N time instance(s) in one reporting instance***   + ***FFS: values of N (e.g., fixed or variable)*** * ***FFS: How to reduce the overhead*** * ***~~Note: The performance gains should be justified given if the increased UCI payload overhead~~*** |
| MediaTek | We support to keep the note as it is not clear the benefit of reporting more than 4 predicted beams, and how to reduce the overhead is still an FFS. |
| InterDigital | Fine |
| Futurewei | OK to us. |
| Qualcomm | We were not initially OK with the proposal (given we do not have enough evidence from 9.2.3.1 about usefulness of the proposal, *given increased overhead*), but for the sake of progress we decided to agree with the inclusion of the suggested note as a compromise. We support the proposal if the note is included. |
| ZTE | For BM-Case2, the prediction results of all N future time instances can be reported in one reporting instance, where the number of reported beams may be larger than 4. However, for each one of future time instance, the benefit of reporting more than 4 beams is not foresee.  Additionally, the associated RSRP in the proposal is not clear. We agree with Panasonic that the associated L1-RSRP could be predicted or measured RSRP. The beam ID to be reported can be based on the AI/ML model inference. However, if both the predicted RSRP and measured RSRP to the same beam are available at the UE side, the measured RSRP should be reported due to its higher reliability. Therefore, we suggest to add a note to clarify this point as Panasonic mentioned.   * ***Note: The associated L1-RSRP could be predicted or measured L1-RSRP.***   Mod: added |
| CATT | We are ok with the updated proposal. |
| Lenovo | We share similar view with Panasonic and ZTE that the associated L1-RSRP may be predicted L1-RSRP or measured L1-RSRP. ZTE’s note is necessary.  Mod: added |
| Xiaomi | We understand that the main point of this proposal is to increase the number of beams in one report instance for UE side model. But we are not clear about the motivation. And we think it is not the enhancement related to AI based beam management. It tries to solve the problem of legacy beam management mechanism. |
| Sony | We are also confused that why need report more than 4 predicted beams for UE-side model. Considering monitoring, it’s understandable, but we think it should be another issue. Waiting more useful information from other companies. |
| LG | Same question as previous round (echoed by multiple companies, Xiaomi/Fujitsu/QC/ZTE/Sony) but no one gives us clear answer.  We may be fine if the proposal changes such that ‘**For BM-Case1 and BM-Case2 with a ~~UE~~NW-side AI/ML model, study potential specification impact ~~(including necessity)~~ of AI model inference from the following additional aspects on top of previous agreements:**’ |
| Ericsson | Don’t understand the update from LG. This is for UE-sided models. Prefer to keep the necessity, it should be very scenario dependent in our view. |
| Mod | @ZTE, Xiaomi, Sony, LGE  For the reporting of Top-K beams, several companies think K<=4 is not sufficient and they show K>4 is beneficial in some evaluations. Thus, they propose to report more than 4 predicted beams.  @Panasonic, ZTE  Note 2 is added |
| CMCC | Fine with this proposal. |
| HW/HiSi | One comment on the Note1: Since we have FFS how to reduce the overhead, it is not for sure that the total overhead will be increased. With increase, is it compared to 4 beams? For example with adaptive number of Top-K beams, the overall overhead would even become smaller.  Therefore we suggest to change the text in the bullet to:   * ***Note1: The performance gains should be justified if ~~given~~ the ~~increased~~ UCI payload overhead is increased*** |
| LG | Thanks FL for further explanation but it is still not very clear. The performance may be improved via increasing # of beams even for legacy beam report but it shall sacrifice UCI overhead. Still unclear what is the key motivation that we need to increase # of beams for UE-predicted beam report compared to legacy beam report. This looks like legacy beam report enhancement, i.e. no relevancy with UE-side AI/ML operation. More explanation would be very helpful in order to study this aspect. Without knowing motivation, we have no idea how to study this. |
| Mod | * Note1 is modified * “***study the necessity, benefit(s), and*** ” is added in main bullet. Hope it can address some concerns. |
| ZTE | Share similar view with LG that the benefit of reporting more than 4 beams for each future time instance is not foresee. According to the current simulation results, reporting Top-2 or Top-4 beams achieves satisfactory beam prediction performance.  We assume that N time instance(s) in the proposal mean the time duration of the prediction window. It's better to clarify the meaning of N. Additionally, according to the proposal, are the prediction results of all N future time instances reported in one reporting instance or separately in different reporting instances? |
| Ericsson | We think it could make sense to report more than 4 beams, considering when having 256 TX beams and 8 RX beams for example. |
| Xiaomi | Thanks for more clarification. Now we can understand the motivation is that the beam prediction accuracy with Top-4 beam may not good enough. But we don’t know the requirement now, is it required to be more than 90% or 99%? In addition, in order to improve the beam prediction accuracy, there are some other solutions such as with more beams in set B. |
| HW/HiSi | Support the proposal (U4)  @ LG, ZTE: We see there is a need to report more than 4 beams, this has been justified by the evaluation where top-K >4 show performance benefits, this is the case for both BM Case 1 and BM Case 2. Since to Top-K is predicted at the UE, and performance gains for K > 4 have been demonstrated by evaluation, we should study how to support it. |
| Lenovo | Although the motivation to report more than 4 beams in a beam report for UE-side model inference it not clear, we can accept the updated version. |
| Fujitsu | Thanks for clarification. We can understand the reason for report more than 4 beams. but from our understanding, if the top-4 beam predication accuracy is not good enough, the performance of AI model needs to be improved (e.g., more advanced model, more beams in set B). This is an implementation issue on UE side. And we are sure what benefits are for gNB with more than 4 inaccurate predicted results if prediction accuracy with top-4 beam is not good enough. |
| Nokia/NSB | OK |
| MediaTek | OK with the proposal |
| Qualcomm | We reviewed the responses from ZTE and Panasonic, but we are not still convinced on the need for Note 2. The main intention of this proposal seems to be reporting more than 4 *predicted* beams. So, adding this note seems to diverge from the main intent. Additionally, reporting more than 4 *measured* beams is already a working assumption from RAN1#110bis-e:  ***Working Assumption***  ***For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the following L1 beam reporting enhancement for AI/ML model inference***   * ***UE to report the measurement results of more than 4 beams in one reporting instance*** * ***Other L1 reporting enhancements can be considered***   So, adding the note seems to be a duplicated effort. |
| Futurewei | Support in principle.  Note there can be two possibilities now based on Note 2 of this proposal.   1. Predicted beams + predicted L1-RSRP; 2. Predicted beams + measured L1-RSRP;   We are not sure of the implication of this but it can be studied. We may want to remove one of them at the end. |
| Mod | As QC/FW comment, the note may lead to more discussion/confusion. We can discuss later how UE select the reported L1-RSRP, e.g., predicted or measured). Thus the note is removed. |
| LG | Thanks for further clarification. But we have similar feeling as Xiaomi/Fujitsu. If the main motivation is for when UE prediction accuracy is not good enough, the whole report may not be useful. Since the proposal is for studying its necessity, the technical motivation should be clear, e.g.  ***Proposal 3.3.2: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, for when the UE-side AI/ML model’s prediction accuracy is not good enough, study the necessity and benefit(s)~~, and potential specification impact~~  of L1 reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) for each time instance. ~~AI model inference from the following additional aspects on top of previous agreements:~~***   * ***~~For BM-Case1: L1 reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) in one reporting instance~~*** * ***~~For BM-Case2: L1 Reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) for each one of N time instance(s) in one reporting instance~~***   + ***~~FFS: values of N (e.g., fixed or variable)~~*** * ***~~FFS: How to reduce the overhead~~*** * ***Note~~1~~: The performance gains should be justified by considering UCI payload overhead***   Mod: In the proposal, “necessity” is included, which indicates the study including whether it is needed, what condition it is needed. Thus, in my understanding, your version (especially the condition) looks like an output of this proposal in some sense.  @Qualcomm, the working assumption you captured is for NW-side model. We haven’t discussed more than 4 beam report for UE-side model. |
| Samsung3 | With the latest updated proposal, it is better to add ‘study the feasibility’.  Mod: would you like to elaborate a bit more why it is not feasibility? |
| Ericsson | Support |
| HW/HiSi | Ok |
| New H3C | OK |
| Spreadtrum | OK with the proposal |
| MediaTek | We are fine with the current proposal |
| Futurewei | Support |
| ZTE | General fine with the proposal if it is the majority view for reporting more than 4 predicted beams. However,  1) for Note2, there is no harm to keep it since the associated L1-RSRP in the proposal is not clear whether it is a predicted or measured L1-RSRP.  2) for BM-Case2, we assume that N time instances in the proposal represent the N future time instances to be predicted. If the intention of this proposal is that in some particular cases, reporting more than 4 predicted beams is needed, which may depend on the confidence level of the predicted beams. Then, it should not be mandatory that the reporting of more than 4 predicted beams is needed for every future time instance. We suggest the following wording change to make it more generic.  ***Proposal 3.3.2: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the necessity, benefit(s), and potential specification impact of AI model inference from the following additional aspects on top of previous agreements:***   * ***For BM-Case1: L1 reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) in one reporting instance*** * ***For BM-Case2: L1 Reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) for ~~each~~ at least one of N time instance(s) in one reporting instance***   + ***FFS: values of N (e.g., fixed or variable)*** * ***FFS: How to reduce the overhead*** * ***Note1: The performance gains should be justified by considering UCI payload overhead*** * ***Note2: The associated L1-RSRP could be predicted or measured L1-RSRP.*** |
| CMCC | Support. |

###### DP 3.3.1

There are also some proposals for other issues, e.g., reporting of predicted beam failure, predicted beam applicable time, how to select the top-K beams, and so on. Since there are usually one or two companies discussing each issue, we can wait for more concrete inputs/proposals from other companies for each of these issues.

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Spec impact of Model monitoring

## General aspects

In previous RAN1 meeting(s), the agreement(s)/conclusion(s) were made as below:

|  |
| --- |
| **RAN1#110**  Agreement  Regarding the model monitoring for BM-Case1 and BM-Case2, to investigate specification impacts from the following aspects   * Performance metric(s) * Benchmark/reference for the performance comparison * Signaling/configuration/measurement/report for model monitoring, e.g., signaling aspects related to assistance information (if supported), Reference signals * Other aspect(s) is not precluded   **RAN1#110bis-e**  Agreement  Study AI/ML model monitoring for at least the following purposes: model activation, deactivation, selection, switching, fallback, and update (including re-training).  FFS: Model selection refers to the selection of an AI/ML model among models for the same functionality. (Exact terminology to be discussed/defined)  Agreement  Study at least the following metrics/methods for AI/ML model monitoring in lifecycle management per use case:   * Monitoring based on inference accuracy, including metrics related to intermediate KPIs * Monitoring based on system performance, including metrics related to system ignalling KPIs * Other monitoring solutions, at least following 2 options.   + Monitoring based on data distribution     - Input-based: e.g., Monitoring the validity of the AI/ML input, e.g., out-of-distribution detection, drift detection of input data, or ~~something simple like checking~~ SNR, delay spread, etc.     - Output-based: e.g., drift detection of output data   + Monitoring based on applicable condition   Note: Model monitoring metric calculation may be done at NW or UE  Agreement  Study performance monitoring approaches, considering the following model monitoring KPIs as general guidance   * Accuracy and relevance (i.e., how well does the given monitoring metric/methods reflect the model and system performance) * Overhead (e.g., ignalling overhead associated with model monitoring) * Complexity (e.g., computation and memory cost for model monitoring) * Latency (i.e., timeliness of monitoring result, from model failure to action, given the purpose of model monitoring) * FFS: Power consumption * Other KPIs are not precluded.   Note: Relevant KPIs may vary across different model monitoring approaches.  FFS: Discussion of KPIs for other LCM procedures  Agreement (AI 9.2.1)  Study at least the following metrics/methods for AI/ML model monitoring in lifecycle management per use case:   * + 1. Monitoring based on inference accuracy, including metrics related to intermediate KPIs     2. Monitoring based on system performance, including metrics related to system ignaling KPIs     3. Other monitoring solutions, at least following 2 options.        - Monitoring based on data distribution          1. Input-based: e.g., Monitoring the validity of the AI/ML input, e.g., out-of-distribution detection, drift detection of input data, or ~~something simple like checking~~ SNR, delay spread, etc.          2. Output-based: e.g., drift detection of output data        - Monitoring based on applicable condition   Note: Model monitoring metric calculation may be done at NW or UE  Agreement (AI 9.2.1)  Study performance monitoring approaches, considering the following model monitoring KPIs as general guidance   * + 1. Accuracy and relevance (i.e., how well does the given monitoring metric/methods reflect the model and system performance)     2. Overhead (e.g., ignaling overhead associated with model monitoring)     3. Complexity (e.g., computation and memory cost for model monitoring)     4. Latency (i.e., timeliness of monitoring result, from model failure to action, given the purpose of model monitoring)     5. FFS: Power consumption     6. Other KPIs are not precluded.   Note: Relevant KPIs may vary across different model monitoring approaches.  FFS: Discussion of KPIs for other LCM procedures  **RAN1#112**  Agreement  Regarding the performance metric(s) of AI/ML model monitoring for BM-Case1 and BM-Case2, study the following alternatives (including feasibility/necessity) with potential down-selection:   * Alt.1: Beam prediction accuracy related KPIs, e.g., Top-K/1 beam prediction accuracy * Alt.2: Link quality related KPIs, e.g., throughput, L1-RSRP, L1-SINR, hypothetical BLER * Alt.3: Performance metric based on input/output data distribution of AI/ML * Alt.4: The L1-RSRP difference evaluated by comparing measured RSRP and predicted RSRP * Other alternatives are not precluded * Note: At least the performance and spec impact should be considered |

The related proposals/ observations are copied as below:

|  |  |
| --- | --- |
| Huawei[2] | *Proposal 22: The input or output data based monitoring (Alt.3), before being further discussed at 9.2.3.2, should be evaluated at 9.2.3.1, including: what metrics can be adopted for evaluating the distribution, how to generate the distribution of data, how accurate the data drift reflects the AI/ML model performance.*  *Proposal 23: For performance metrics of AI/ML model monitoring under BM-Case1 and BM-Case2, study the following alternatives as a starting point:*   * *Alt.1: Beam prediction accuracy related KPIs, e.g. accuracy of predicted beam ID and/or predicted RSRP.* * *Alt.2: Link quality related KPIs, e.g. throughput, L1-RSRP, L1-SINR, hypothetical BLER, etc.*   *Proposal 24: For AI/ML model monitoring for BM-Case1 and BM-Case2, study at least the following benchmarks for performance comparison:*   * *Benchmark 1 (upper bound): The genie-aided best beam(s)/RSRP(s) obtained by measuring beams of Set A.* * *Benchmark 2 (lower bound): The determined beam under an alternative BM solution of comparable beam sweeping overhead/latency with the undergoing model. Two cases can be studied for this alternative BM solution:*   + *Case 1: Non-AI/ML solution, to make the decision of deactivation/fallback based on the performance comparison with the AI/ML solution being monitored.*   + *Case 2: AI/ML solution subject to an inactive model, to make the decision of switching/selection based on the performance comparison with the AI/ML solution being monitored.*   *Proposal 25: For AI/ML model monitoring for BM-Case1 and BM-Case2, to enable the performance comparison between the undergoing AI/ML solution and an alternative BM solution (e.g., non-AI/ML or an inactive AI/ML model), the following options can be studied for the UE report:*   * *Option 1: Separately report the performance metrics for the undergoing AI/ML solution and the alternative BM solution.* * *Option 2: Report the relative gap between the undergoing AI/ML solution and the alternative BM solution.*   *Proposal 26: For AI/ML model monitoring for BM-Case1 and BM-Case2, to minimize the sweeping overhead for the performance comparison between the active AI/ML solution and an alternative BM solution (e.g., non-AI/ML or an inactive AI/ML model), study the approach to enable the performance measurement of the undergoing AI/ML solution, upper bound benchmark, and lower bound benchmark within one beam sweeping occasion.* |
| ZTE[4] | *Proposal 27: All alternatives of performance metrics for AI/ML model monitoring should be evaluated in agenda 9.2.3.1 before further down-selection.*  *Proposal 28: Prioritize beam prediction accuracy related KPIs (i.e., Alt.1 and Alt.4) as the performance metric for AI/ML model monitoring since it has been evaluated in agenda 9.2.3.1 and could well reflect the performance of the AI/ML model.*  *Observation 13: Considering ping pong effect in wireless communication, the performance of the AI/ML model may change with time and inference errors occur even for a valid AI/ML model.*  *Proposal 29: Study performance monitoring mechanisms to claim that an AI/ML model/functionality is no longer valid, e.g., AI/ML inference fails for several consecutive times or the probability of inference failure exceeds a certain threshold.*  *Observation 14: The selection of benchmark for performance comparison has a strong correlation with the performance metric of AI/ML model monitoring.*  *Proposal 30: Support to take Alt.1 (i.e., the best beam(s) obtained by measuring beams of a set indicated by gNB) as the benchmark/reference for performance comparison, which corresponds to an upper bound of beam prediction.* |
| Vivo[5] | *Proposal 38: Regarding the performance metric(s) of AI/ML model monitoring for BM-Case1 and BM-Case2, support Alt.1 and Alt.4,*  * Alt.1: Beam prediction accuracy related KPIs, e.g., Top-K/1 beam prediction accuracy*  * Alt.4: The L1-RSRP difference evaluated by comparing measured RSRP and predicted RSRP*  *Proposal 39: Regarding the performance metric(s) of AI/ML model monitoring for BM-Case1 and BM-Case2, deprioritized Alt.3, i.e. performance metric based on input/output data distribution of AI/ML*  *Proposal 40: Support Alt.1, i.e., the best beam(s) obtained by measuring beams of a set indicated by gNB, as the benchmark/reference for performance comparison for AI/ML model monitoring for BM-Case1 and BM-Case2.* |
| OPPO[6] | *Proposal 15: For performance metric of AI/ML model monitoring, further study the beam prediction accuracy related KPIs.* |
| Spreadtrum[7] | *Observation 2: Considering the reference for the performance comparison,*   * *If set A is used as the reference, UE reporting overhead may be significant.*   *- If set B is used as the reference, only part of the output results will be compared.*  *Proposal 10: For BM-Case1 and BM-Case2, actual RSRP of Set A/ Set B used as the reference for the performance comparison can be considered.*  *Proposal 11: For BM-Case1 and BM-Case2, the RSRP difference evaluated by comparing actual RSRP and predicted RSRP can be used as a performance metric.* |
| CATT[9] | *Proposal 15: Regarding the model monitoring for BM-Case1 and BM-Case2, study the following aspects for the intermediate KPI calculation:*   * *How to calculate the intermediate KPIs based on both of Set A and Set B;* * *How to calculate the intermediate KPIs only based on Set B.* |
| IDC[11] | *Proposal 14: For AI/ML monitoring, consider a common mechanism for multiple purposes, procedures for identifying need of AI/ML model recovery, UE request/gNB trigger and AI/ML model recovery.*  *Observation 11: Supporting multiple KPIs can be beneficial as each KPI has different functionalities.*  *Proposal 15: A procedure to update AI/ML model and corresponding AI/ML model input based on UE monitoring of multiple performance metrics should be considered.* |
| Sony[12] | *Proposal 5 : Support Alt.2 and Alt.4 as the performance metric(s) of AI/ML model monitoring for BM-Case1 and BM-Case2.*  *Proposal 6 : Multi times output of AI/ML should be monitored. And the monitoring interval which is represented by the times of AI/ML output depends on channel fading degree.*  *Proposal 7 : Measurement report with AI beam management procedure shall also be collected to build a model for estimating the RSRP of a given scenario including UE mobility, time of operation, time scale of prediction, etc.*  *Proposal 8 : Measurement report with traditional beam management procedure shall also be collected to build a model for estimating the RSRP of a given scenario including UE mobility, etc.*  *Proposal 9 : Need to compare the RSRP of AI based beam management with the RSRP achieved with traditional RSRP at the same environment* |
| Ericsson[14] | **Table 1** **Summary of different performance metrics-based methods for AI/ML BM model monitoring**   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Performance metric** | **Examples** | **Required data samples for to derive the performance metric** | **Benefits** | **Challenges** | | **Inference Accuracy**  (Intermediate KPIs)  (Alt.1, Alt. 4) | Beam prediction accuracy related KPIs, e.g., Top-K/1 beam prediction accuracy | The measured L1-RSRP values of the predicted top-K/1 beam(s) + the ground truth (i.e., the measured RSRP value of the best beam in set A) | Metric reflects the model performance very well  Expected to provide reliable model failure detection | Signalling overhead for collecting ground truth data at UE/NW (RS transmission and/or UE reporting)  Frequent monitoring degrades the usability of the model.  May not reflect the system performance very well (e.g., a higher prediction accuracy does not necessarily mean a better system KPI) | | The L1-RSRP difference evaluated by comparing measured RSRP and predicted RSRP | The predicted L1-RSRP values of the predicted top-K/1 beam(s) + the ground truth (i.e., the measured L1-RSRP values of these beams) | | **System/Link performance metric(s)**  (Alt.2) | Throughput | Throughput values using AI/ML model. Reference throughput values for a non-AI/ML solution. | Metric reflects the system performance  Low complexity and signalling overhead | Challenging to identify that the degradation is due to an inaccurate model (inaccurate model monitoring) | | Number of beam failures or/and beam switches within a time window | Number of beam failure instances, number of triggered beam failure recovery procedures, number of beam switches | | BLER for a hypothetical PDCCH transmission | SSB/CSI-RS beam(s) measurements (used for calculated the BLER of a hypothetical PDCCH transmission at the PHY layer of UE, measurements can be performed on serving beams or configured/dedicated beams); a target BLER (configured by e.g., the network) + hypothetical PDCCH configuration (configured by e.g., the network, or defined in spec) | | **Data distribution**  (Alt.3) | Input/output data distribution of AI/ML | The input data can be the measured RSRP values for Set B, and the output data can be the AI/ML model output. | No additional signalling overhead for obtaining input/output data  Shorter latency for obtaining data samples for model monitoring  Frequent monitoring possible | May not reflect model performance as well as Alt.1  May not reflect system performance as well as Alt. 2  To achieve reliable model failure detection, many samples may be required to calculate statistical metrics. This may lead to   * Potential high complexity (computation and memory cost) * Potential long monitoring window, hence, increased latency from model failure occurs to detecting the failure | |
| Xiaomi[16] | *Proposal 13: gNB to transmit all beams in set A periodically/semi-persistently/ a-periodically for performance monitoring.*  *Proposal 17: Alt.1 with Beam prediction accuracy related KPIs and Alt. 4 with predicted L1-RSRP difference can be used as metric for performance monitoring.* |
| Google[17] | *Proposal 7: For spatial domain beam prediction, the beam quality for current beam from an indicated TCI can be used for performance validation, and if none of the predicted beam(s) can provide better beam quality than current beam, the predicted beam(s) are assumed to fall to pass the performance validation.*  *Proposal 12: For time-domain beam prediction, the beam quality for current beam from an indicated TCI can be used for performance validation, and if none of the predicted beam(s) can provide better beam quality than current beam, the predicted beam(s) are assumed to fall to pass the performance validation.*  *Proposal 13: Study UE feedback before the beam action time for performance validation for predicted beam in addition to the ACK/NACK for the TCI update signaling.* |
| Samsung[19] | *Proposal 11. For the performance metric(s) of AI/ML model monitoring, the necessity and feasibility of Alt-3 (i.e., performance metric based on input/output data distribution of AI/ML) should be further studied.* |
| ETRI[21] | *Proposal 1: Regarding the performance metric(s) of AI/ML model monitoring for BM-Case1 and BM-Case2, support to study the following as a priority:*  *- Alt.1: Beam prediction accuracy related KPIs, e.g., Top-K/1 beam prediction accuracy*  *- Alt.2: Link quality related KPIs, e.g., throughput, L1-RSRP, L1-SINR, hypothetical BLER* |
| CMCC[22] | *Proposal 9: For model inference of BM-Case1, beam prediction accuracy related KPI can be used as the metric of model performance monitoring.* |
| MediaTek[23] | *Proposal 4: We support proposal 5.1.2 regarding the performance comparison benchmark/reference for AI/ML model monitoring of BM-Case1 and BM-Case2 with the following updates:*  *For AI/ML model monitoring for BM-Case1 and BM-Case2, study the following alternatives the benchmark/reference (if applicable) for performance comparison as a starting point:*   * *Alt.1: The best beam(s) obtained by measuring beams of a set indicated by gNB (e.g., Beams from Set A)* * *Alt.2: The predicted best beam(s) obtained by model output (e.g., Predicted Top-K Beams)* * *FFS: Alt.~~2~~3: The best beam(s) among those used for AI/ML model inputs (e.g., Beams of Set B)* * *FFS: Alt.~~3~~4: The beam corresponding to some indicated TCI state(s)* * *Other alternatives are not precluded* * *Note1: the performance and spec impacts should be considered* * *Note2: Legacy mechanism may be reused* |
| NVIDIA[24] | *Proposal 9: For AI/ML based beam prediction in spatial/time domain, study potential specification impact related to assistance signalling and procedure for model performance monitoring and model update/tuning.* |
| Lenovo[26] | *Proposal 9: Select Alt 1 and Alt 4 as the performance metric(s) of AI/ML model monitoring.*  * Alt.1: Beam prediction accuracy related KPIs, e.g., Top-K/1 beam prediction accuracy*  * Alt.4: The L1-RSRP difference evaluated by comparing measured RSRP and predicted RSRP* |
| DOCOMO[29] | *Observation 2: Beam prediction accuracy related KPI requires the large measurement overhead for searching out the actual top 1/K beam(s).*  *Observation 3: Link quality related KPI is affected by various factors other than model performance.*  *Proposal 5: Discuss the feasibility of the performance monitoring based on the input/output data distribution in the beam prediction, before the specification impact discussion related to it.*  *Observation 4: The required measurement overhead for performance monitoring is relatively low for the monitoring based on the predicted L1-RSRP difference.* |

###### Proposal 4.1.1

For the benchmark/reference, there was a proposal discussed in previous meetings, but no consensus was achieved. In this meeting, several tdocs continue to discuss this issue and more alternative were proposed. A proposal is provided with various alternatives, and collect companies’ views on these alternatives.

The related proposals in tdocs are as below:

* Huawei: Proposal 24
* Vivo: Proposal 30
* Spreadtrum: Proposal 10
* Sony: Proposal 9
* Google: Proposal 7, 12
* MTK: Proposal 14

The following proposal were discussed in last meeting. Based on the submitted tdocs and previous discussions, most companies support Alt.1 as the start point and doubt the motivation/benefits to introduce Alt.2 and Alt.3. Thus, FFS is put for Alt.2 and Alt.3.

***Proposal 4.1.1: For AI/ML model monitoring for BM-Case1 and BM-Case2, study potential specification impact of at least the following alternatives as the benchmark/reference (if applicable) for performance comparison:***

* ***Alt.1: The best beam(s) obtained by measuring beams of a set indicated by gNB (e.g., Beams from Set A)***
  + ***FFS: gNB configures one or multiple sets for one or multiple benchmarks/references***
  + ***FFS: the definition of “best beam(s)”***
* ***FFS:***
  + ***Alt.3: The beam corresponding to some or all the indicated/activated TCI state(s)***
  + ***Alt.4: Measurements of the predicted best beam(s) obtained by model output (e.g., Comparison between actual L1-RSRP and predicted RSRP of predicted Top-1/K Beams)***
  + ***~~Alt.6: AI/ML solution subject to an inactive model, to make the decision of switching/selection based on the performance comparison with the AI/ML solution being monitored.~~***

|  |  |  |
| --- | --- | --- |
|  | Support | Not support |
| Alt.1 | Nokia/NSB, New H3C, DCM, vivo, Spreadtrum, CAICT, CATT,CMCC, Ericsson, Xiaomi, Samsung, Sony, Fujitsu, InterDigital, NVIDIA, QC, MediaTek, HW/HiSI, ZTE |  |
| Alt.2 |  | Nokia/NSB |
| Alt.3 | Nokia/NSB (not clear as an alternative with current wording) |  |
| Alt.4 | DCM,CATT, MediaTek, vivo | Nokia/NSB (not clear as an alternative with current wording) |
| Alt.5 | CATT,CMCC, Xiaomi, Sony, HW/HiSi | Nokia/NSB (this is more or less Alt.1) |
| Alt.6 | DCM,CATT, Xiaomi, HW/HiSi,CMCC | Nokia/NSB |

|  |  |
| --- | --- |
| Company | Comments |
| Mod | Companies are also invited to fill the above table. The current proposal will be refined (e.g., remove some alternatives”) based on the inputs. |
| Nokia/NSB | There are clear agreements in 9.2.3.1 on these. Not sure baseline can be anything than measuring the ground truth of Set A. |
| Google | Support Alt3 only. But we suggest changing Alt3 as follows:   * ***Alt.3: The beam corresponding to some or all the indicated/activated TCI state(s)*** |
| New H3C | Only support Alt.1 |
| NTT DOCOMO | The required beam measurement/reference is different according to the performance metric. For example, if the beam prediction accuracy is used as a performance metric, the best beam from Set A and the predicted best beam are necessary. Likewise, if the predicted L1-RSRP difference is the performance metric, the measurement of the predicted best beam and predicted values of the predicted best beam are necessary. Thus, it should be discussed per performance metric.  Mod: Since there are 4 alternatives for performance metric, it seems impossible to discuss all the alternatives for each metric. Here, we can select some options supported by majority companies and the FFS list is still kept. In my understanding, for your 2nd case (i.e., the predicted L1-RSRP difference is the performance metric), Alt.5 can work since the measured L1-RSPR can obtained by legacy beam sweeping. |
| Vivo | We support Alt 1 with the set to be set A. |
| HW/HiSi | Support in principle  We think that also multiple benchmarks could be supported and a corresponding note could added.  For Alt2, why input is it based on model input, can the idea behind it please be explained?  Alternative 3 needs some more clarification, sn’t’t this also a non-AI approach and covered by Alt5? |
| CAICT | Alt 1 is preferred. |
| CATT | We prefer to support Alt.1 with modification and Alt 4. We are open to discuss Alt.5 and Alt.6.  For the benchmark/reference for performance comparison for model monitoring, one feasible method is that UE measures multiple Set Bs to constitute Set A or nearly constitute Set A, so that UE can get the best beam(s) of Set A based on the measurement of multiple Set Bs. Hence, we suggest the following update:  ***Proposal 4.1.1: For AI/ML model monitoring for BM-Case1 and BM-Case2, study the following alternatives as the benchmark/reference (if applicable) for performance comparison as a starting point:***   * ***Alt.1: The best beam(s) obtained by measuring beams of ~~a~~ set(s) indicated by gNB (e.g., Beams from Set A)*** * ***Alt.2: The best beam(s) among those used for AI/ML model inputs (e.g., Beams of Set B)*** * ***Alt.3: The beam corresponding to some indicated TCI state(s)*** * ***Alt.4: The predicted best beam(s) obtained by model output (e.g., Predicted Top-K Beams)*** * ***Alt.5: Non-AI/ML solution, to make the decision of deactivation/fallback based on the performance comparison with the AI/ML solution being monitored.*** * ***Alt.6: AI/ML solution subject to an inactive model, to make the decision of switching/selection based on the performance comparison with the AI/ML solution being monitored.***   Mod: it seems not a normal solution. Let’s hear more views. |
| CMCC | The difference between Alt2 and Alt5 is not clear. |
| Ericsson | Prefer alternative 1. We can be ok with 2 and 5. Alt 3 and 6 is too vague and needs clarification. |
| Xiaomi | We support Alt 1 and open to discuss with Alt 5&6. For Alt 3, it is not clear how to obtain the indicated TCI state(s). for Alt 4, is it to compare predicted best beam with predicted best beam? |
| Samsung | We are open to study all Alts as a start point. In our understanding, we are not down selecting one alternative in this meeting, so motivation of the above table is unclear. |
| Sony | In our understanding, we can periodically compare the performance between ***Non-AI/ML solution*** and ***AI/ML solution***, and then make the decision whether to perform fallback or not. So, we support Alt.5 as well. |
| Fujitsu | Support Alt.1. |
| Qualcomm | The intention and the corresponding specification impact of the proposal is not clear. For the purpose of evaluations in 9.2.3.1, there are some prior agreements to coordinate results across companies, but it is not clear what are the spec implications of the proposal. The natural choice would be to consider measuring the ground truth from Set A, but not sure why we need to list all these alternatives and what are the corresponding spec implications.  Mod: In order to assess the active AI model is good or not, the UE/NW may need to compare it performance with something (we refer it as benchmark/reference). |
| Lenovo | We support Alt.1 |
| MediaTek | In our opinion, Alt.4 is a more useful candidate than Alt.2. As Alt.4 has higher possibility to use a beam with high value of L1-RSRP for benchmarking than Alt.2, where the best beam in Set B might still have very low L1-RSRP values. For example, if we use predicted L1-RSRP error as the performance metrics, it is more meaningful to monitor the predicted L1-RSRP error on the benchmark beams with high L1-RSRP values than on beams with low L1-RSRP values. |
| Mod | It seems Alt.1 is supported by majority companies. |
| ZTE | Alt.1 can be a baseline as it corresponds to an upper bound of beam prediction and can be easily obtained by a legacy beam sweeping procedure. Besides, it can directly reflect how well the AI/ML model is working. |
| NEC | We prefer Alt.1 and Alt.6. |
| HW/HiSi | We prefer Alt.1, Alt.5 and Alt 6.  Are fine the current proposal for progress. It is our understanding that Alt5 also can be covered by Alt.1  Mod: Alt.5 is copied from HW’s tdoc. Thus, if it is covered by Alt.1, then I will delete it from the list |
| InterDigital | We are fine with Alt 1. For other alternatives, we are not sure what are indented operations as there are not enough details. |
| Futurewei | We support Alt.1 and Alt.5. |
| Lenovo | We support Alt1. |
| Xiaomi | We are fine with the updated proposal |
| Sony | We prefer Alt.1 and Alt.5 |
| Nokia/NSB | We do not understand the Alt.4. Is the idea to measure Top-K? If yes, then it is related to measuring part of Set A.  Alt.5 also not clear. What is the non-ML solution?  Alt. 6, monitoring based on inactive model is being discussed in 9.2.1. We do not fully get that as well. We can have other options are not precluded.  ***Proposal 4.1.1: For AI/ML ~~model~~ monitoring for BM-Case1 and BM-Case2, at least study the following alternatives as the benchmark/reference (if applicable) for performance comparison as a starting point:***   * ***Alt.1: The best beam(s) obtained by measuring beams of a set indicated by gNB (e.g., Beams from Set A)*** * ***FFS:***   + ***Alt.2: The best beam(s) among those used for AI/ML model inputs (e.g., Beams of Set B)***   + ***Alt.3: The beam corresponding to some or all the indicated/activated TCI state(s)***   + ***Other options are not precluded***   + ***~~Alt.4: The predicted best beam(s) obtained by model output (e.g., Predicted Top-K Beams)~~***   + ***~~Alt.5: Non-AI/ML solution, to make the decision of deactivation/fallback based on the performance comparison with the AI/ML solution being monitored.~~***   + ***~~Alt.6: AI/ML solution subject to an inactive model, to make the decision of switching/selection based on the performance comparison with the AI/ML solution being monitored.~~*** |
| NTT DOCOMO | We do not quite get the intention of this proposal. What is motivation to agree the benchmark/reference for performance comparison? If it is for the evaluation purposes, it should be discussed in 9.2.2.1. If it is for the potential spec impact study of the signaling for the monitoring entity to obtain the benchmark/reference, that should be mentioned.  Mod: The main bullet is modified. Please check whether it is clear or not |
| HW/HiSi | Multiple benchmarks have been addressed during the last meeting, e.g. upper bound compared to best beam from Set A , lower bound according to a conventional beam indication, beam from another non-active AI model output. For the course of this discussion these different benchmarks should be kept. Depending on the strategy how to obtain the best beam(s) in with Alt1, these different benchmarks can be realized. To clarify that we suggest to add the following note to Alt1:  ***Proposal 4.1.1: For AI/ML model monitoring for BM-Case1 and BM-Case2, at least study the following alternatives as the benchmark/reference (if applicable) for performance comparison as a starting point:***   * ***Alt.1: The best beam(s) obtained by measuring beams of a set indicated by gNB (e.g., Beams from Set A)***   + ***Note: The best beam(s) can be obtained from one or multiple policies (e.g. best global beam from Set A, best beams indicated according to non-AI solution, best beam according to a non-active AI model, etc.)*** * ***FFS:***   + ***Alt.2: The best beam(s) among those used for AI/ML model inputs (e.g., Beams of Set B)***   + ***Alt.3: The beam corresponding to some or all the indicated/activated TCI state(s)***   + ***Alt.4: The predicted best beam(s) obtained by model output (e.g., Predicted Top-K Beams)***   + ***Alt.5: Non-AI/ML solution, to make the decision of deactivation/fallback based on the performance comparison with the AI/ML solution being monitored.***   + ***Alt.6: AI/ML solution subject to an inactive model, to make the decision of switching/selection based on the performance comparison with the AI/ML solution being monitored.*** |
| Mod | * Main bullet is modified according to DCM’s comments * Alt.2 is removed. It is the low bound. Thus, if the predicted beam is better than all the beams of Set B, it seems difficult to declare the prediction performance is good. The proponents please feel free correct me if I missed something. * Alt.5 is deleted. Alt.5 is copied from HW’s tdoc and HW commented it is covered by Alt.1   @Proponents of Alt.4 (e.g., MTK, …): If I understood correctly, the entity (UE/NW) will compare the predicted L1-RSRP and the measurement L1-RSRP. Thus, the reference should be the measured L1-RSRP which is obtained by beam sweeping. For UE measurement, there will be a set configured by gNB. In this case, it seems the reference is Alt.1. Please feel free to correct me if I missed something.  @HW: The new note seems to make things more complicated. The current Alt.1 is quite inclusive and doesn’t imply or preclude any scheme to get the information. |
| NTT DOCOMO | We thank FL to explain the motivation of this proposal.  With that in mind, we think Alt.4 should be further prioritized to calculate the predicted L1-RSRP, which does not require the overhead to search out the actual Top-1/K beam(s). To address the Nokia’s concern, we suggest updating as follows for the clarification.  ***Alt.4: Measurements of the predicted best beam(s) obtained by model output (e.g., Comparison between actual L1-RSRP and predicted RSRP of predicted Top-K Beams)*** |
| ZTE | Open to discuss. A minor comment on top of DCM’s update.  ***Alt.4: Measurements of the predicted best beam(s) obtained by model output (e.g., Comparison between actual L1-RSRP and predicted RSRP of predicted Top-1/K beams)*** |
| CATT | For Alt.5, we are fine to delete it.  For Alt.1, to get the best beam(s) of Set A, gNB still needs to transmit RS with beams in Set A. Thus, the RS overhead may be large. One feasible method is that UE measures multiple sets to constitute Set A or nearly constitute Set A. We Support to add the following alternative:   Alt X: The best beam(s) obtained by measuring sets of beams that nearly constitutes set A. |
| Ericsson | Support Docomo’s update. |
| Xiaomi | For Alt 3, it is not clear how to obtain the indicated TCI state(s). for Alt 4, is it to compare predicted best beam with predicted best beam? |
| HW/HiSi | @FL: “The new note seems to make things more complicated. The current Alt.1 is quite inclusive and doesn’t imply or preclude any scheme to get the information.”  Our intention is of course not to make the proposal more complicated, but to make it clearer. In the previous meetings, multiple benchmarks have been discussed, such as upper bound (best beam from Set A), lower bound for fall-back, inactive AI model, etc. Also it has been discussed whether to have one more multiple benchmark metric.  We think that the note for Alt1 is important since it clarifies that the different benchmarks can be derived and also that different benchmarks can be used together, this is even more important now, since some of the FFS are starting to get removed.  Also, in our understanding all the different benchmarks need an indication of beams from the gNB. Therefore, Alt1 is involved in the different sub-options. The important point is in our understanding is whether one or multiple benchmarks can be used at the same time. In our view it useful to have multiple for different purposes, e.g. model updating, fall-back to non-AI etc,.  By adding the note, we could simplify the proposal and remove the FFSs.  ***Updated Proposal 4.1.1: For AI/ML model monitoring for BM-Case1 and BM-Case2, ~~at least~~ study potential specification impact of at least the following alternatives as the benchmark/reference (if applicable) for performance comparison ~~as a starting point~~:***   * ***Alt.1: The best beam(s) obtained by measuring beams of a set indicated by gNB (e.g., Beams from Set A)***   + ***Note: The best beam(s) can be obtained from one or multiple benchmarks (e.g. best global beam from Set A, best beams indicated according to non-AI solution, best beam according to a non-active AI model, etc.)*** * ***~~FFS:~~***   + ***~~Alt.2: The best beam(s) among those used for AI/ML model inputs (e.g., Beams of Set B)~~***   + ***~~Alt.3: The beam corresponding to some or all the indicated/activated TCI state(s)~~***   + ***~~Alt.4: The predicted best beam(s) obtained by model output (e.g., Predicted Top-K Beams)~~***   + ***~~Alt.5: Non-AI/ML solution, to make the decision of deactivation/fallback based on the performance comparison with the AI/ML solution being monitored.~~***   + ***~~Alt.6: AI/ML solution subject to an inactive model, to make the decision of switching/selection based on the performance comparison with the AI/ML solution being monitored.~~*** |
| Lenovo | In addition to Alt.1, we are fine with DOCOMO’s updated Alt.4. At least for the model can output predicted L1-RSRP, the gap between predicted RSRP and the measured RSRP of the same beam pair can be used for model monitoring. |
| Nokia/NSB | Ok with the clarification from DCM.  We do not support listing Alt.6. Why to bother discussing inactive models as that is not visible to the gNB. Also, how come it is an inactive model if that is used for inference. |
| MediaTek | To answer Nokia’s comment, we think Alt. 1 cannot fully cover Alt.4. The wording of Alt. 1 indicates actions of beam sweeping and then finding some best beam(s) (e.g. Top-K) as reference/benchmark beams. However, for Alt.4, the benchmark beams are not obtained from beam sweeping, but from the model output. After obtaining model output, gNB configures the predicted Top-K beams for UE measurement. Then, the entity (UE/NW) compare the predicted L1-RSRP and the measurement L1-RSRP. We support DCM’s update if it helps other companies to understand the concept. |
| InterDigtal | We prefer to remove all the alternatives under the FFS and further discuss in the next meeting. |
| Mod | The proposal is updated based DCM’s version and ZTE’s modification for Alt.4  @Huawei: Would you like to response to Nokia’s comment on Alt.6? |
| Qualcomm | Fine with the current form of Alt. 4, but do not support Alt. 6 for a similar reason mentioned by Nokia. |
| Futurewei | We can go with HW’s latest revision. But since there is only one alternative left. The wording should be change accordingly. Such as below (on top of HW’s proposal, new changes in purple).  ***Updated Proposal 4.1.1: For AI/ML model monitoring for BM-Case1 and BM-Case2, ~~at least~~ study potential specification impact of ~~at least~~ the following ~~alternatives~~ as the benchmark/reference (if applicable) for performance comparison ~~as a starting point~~:***   * ***~~Alt.1:~~ The best beam(s) obtained by measuring beams of a set indicated by gNB (e.g., Beams from Set A)***   + ***Note: The best beam(s) can be obtained from one or multiple benchmarks (e.g. best global beam from Set A, best beams indicated according to non-AI solution, best beam according to a non-active AI model, etc.)*** |
| Mod | The proposal is updated.   * An FFS part is added in Alt.1 * Alt.6 is removed. * Some other alternatives are kept as several companies prefer them |
| LG | For us, it is very hard to understand how the benchmark/reference for performance comparison impacts specification. Could someone clarify one example?  Mod: Let take your suggested “the event” for example. Assume a benchmark is the best beam of Set A.  The event may be defined as the performance of the current AI model/functionality is worse than this benchmark with some conditions |
| InterDigital | We are still not fine with the alternatives for the FFS part. |
| CATT | 1. Support to add the FFS part in Alt.1, but the e.g. part (e.g., Beams from Set A) can be removed.   Mod: It is just an example. It seems no harm to keep it.   1. We do not support to remove Alt.6. To determine the most suitable AI/ML model in current time, both the activated AI/ML model and the inactive model(s) should be monitored, otherwise NW or UE will not know which model is the most suitable in current time or which model should be switched to if the performance of activated model for inference is not desirable. Therefore, the performance of inactive model(s) should be the benchmark/reference. Regarding the performance metric of monitoring inactive model(s), e.g., based on input data distribution/applicable condition/inference accuracy can be discussed later.   Mod: Let’s hear more views |
| Samsung3 | We can understand the intention of HW. May the following revision can be considered by the group.  ***Updated Proposal 4.1.1: For AI/ML model monitoring for BM-Case1 and BM-Case2, ~~at least~~ study potential specification impact of at least the following alternatives as the benchmark/reference (if applicable) for performance comparison ~~as a starting point~~:***   * ***Alt.1: The best beam(s) obtained by measuring beams of a set indicated by gNB ~~(e.g., Beams from Set A)~~***   + ***FFS: the definition of ‘best’*** |
| Ericsson | Support the updates to Alt1 from Samsung, and FLs update to 4.  We have similar concerns as Xiaomi, for Alt 3, it is not clear how to obtain the indicated TCI state(s). Our preference is to focus on alt 1 and 4.  Mod: It is put in FFS part. Proponent may provide more details in the following studies. Let’s hear more views. |
| Hw/HiSi | Support, and also fine with updates from FW and E///. |
| Mod | The proposal is updated by adding “FFS: the definition of “best beam(s)” as suggested by SS. |
| New H3C | OK |
| vivo | After going through all the discussions, we see the need of Alt 4 as it can help to reduce reporting and RS overhead. Hence we support it as well.  Hence we share similar as Ericsson to focus on Alt 1 and Alt 4.  The following is our suggested revision.  ***Proposal 4.1.1: For AI/ML model monitoring for BM-Case1 and BM-Case2, ~~at least~~ study potential specification impact of at least the following alternatives as the benchmark/reference (if applicable) for performance comparison ~~as a starting point~~:***   * ***Alt.1: The best beam(s) obtained by measuring beams of a set indicated by gNB (e.g., Beams from Set A)***   + ***FFS: gNB configures one or multiple sets for one or multiple benchmarks/references***   + ***FFS: the definition of “best beam(s)”*** * ***Alt.4: Measurements of the predicted best beam(s) obtained by model output (e.g., Comparison between actual L1-RSRP and predicted RSRP of predicted Top-1/K Beams)*** * ***FFS:***   + ***Alt.3: The beam corresponding to some or all the indicated/activated TCI state(s)***   + ***~~Alt.4: Measurements of the predicted best beam(s) obtained by model output (e.g., Comparison between actual L1-RSRP and predicted RSRP of predicted Top-1/K Beams)~~***   + ***~~Alt.6: AI/ML solution subject to an inactive model, to make the decision of switching/selection based on the performance comparison with the AI/ML solution being monitored.~~***   Mod: Let’s check other companies’ views |
| Spreadtrum | We were also a little confused about Alt 3. Is the meaning of "indicated TCI states" similar to “ indicated beams”? If so, beam should be used to describe it more clearly  Mod: It is in FFS part. The wording can be further refined when we decide to support it. |
| MediaTek | Agree with vivo, we can remove FFS for Alt.4. |
| Futurewei | Support |
| ZTE | We support to focus on Alt1 and Alt4 as with Ericsson and vivo. Besides, the first FFS is confusing whether multiple benchmarks/references are obtained at the same time or different benchmarks/references are used at different times. We prefer to make it more clear as below.   * ***Alt.1: The best beam(s) obtained by measuring beams of a set indicated by gNB (e.g., Beams from Set A)***   + ***FFS: gNB configures one or multiple sets for obtaining one or multiple benchmarks/references at the same time***   + ***FFS: the definition of “best beam(s)”*** |
| NTT DOCOMO | Agree with Ericsson, vivo, ZTE, and MediaTek. FFS on Alt.4 should be removed. |
| CMCC | Support Alt1 and Alt4. The motivation of Alt3 is not clear to us, can companies clarify the motivation? |
| HW/HiSi | Ok |

###### DP 4.1.1

In the last meeting, four alternatives were agreed for the further study of performance metric. In the submitted tdocs, many companies show their preference on these alternatives. The related proposals in tdocs are as below:

* Huawei: Proposal 22, 23
* ZTE: Proposal 27, 28
* Vivo: Proposal 38, 39
* OPPO: Proposal 15
* Spreadtrum: Proposal 11
* CATT: Proposal 15
* IDC: Observation 11
* Sony: Proposal 5
* Ericsson: Table 1
* Xiaomi: Proposal 17
* Samsung: Proposal 11
* ETRI: Proposal 1
* CMCC: Proposal 9
* Lenovo: Proposal 9

Companies’ views are summarized as below:

|  |  |  |
| --- | --- | --- |
|  | Support | Deprioritize or need further check |
| Alt.1: Beam prediction accuracy related KPIs | Huawei, ZTE, vivo, OPPO, Xiaomi, ETRI, CMCC, Lenovo, Ericsson, QC |  |
| Alt.2: Link quality related KPIs | Huawei, Sony, ETRI, Ericsson, QC |  |
| Alt.3: Performance metric based on input/output data distribution of AI/ML | Ericsson, QC | Huawei, ZTE, vivo, SS, DCM |
| Alt.4: The L1-RSRP difference evaluated by comparing measured RSRP and predicted RSRP | ZTE, vivo, Spreadtrum, Sony, xiaomi, Lenovo,Ericsson, QC, DCM |  |

Vivo suggests to deprioritize Alt.3 whereas HW/ZTE/SS/DCM suggest to study the feasibility of Alt.3. Considering this is the first meeting since the group made the agreement. We can wait for more evaluation results and then decide further down-select the alternatives.

|  |  |
| --- | --- |
| Company | Comments |
| Ericsson | We have evaluations on Alt.3. |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## NW-side model

In previous RAN1 meeting(s), the agreement(s)/conclusion(s) were made as below:

|  |
| --- |
| **RAN1#110bis-e**  Agreement  For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the NW-side model monitoring:   * NW monitors the performance metric(s) and makes decision(s) of model selection/activation/ deactivation/switching/ fallback operation   Agreement  Regarding NW-side model monitoring for a network-side AI/ML model of BM-Case1 and BM-Case2, study the potential specification impacts from the following aspects   * Beam measurement and report for model monitoring * Note: This may or may not have specification impact.   **RAN1#111**  Agreement  Regarding NW-side model monitoring for a network-side AI/ML model of BM-Case1 and BM-Case2, study the necessity and the potential specification impacts from the following aspects:   * UE reporting of beam measurement(s) based on a set of beams indicated by gNB. * Signaling, e.g., RRC-based, L1-based. * Note: Performance and UE complexity, power consumption should be considered. |

|  |  |
| --- | --- |
| Huawei[2] | *Observation 12: It is necessary to support real time UE reporting for NW-side monitoring to enable fast identification of network performance fluctuation/degradation.*   * *E.g., in case of performance degradation, event-triggered monitoring window can be activated so that gNB can efficiently collect data and thereby quickly identify whether the degradation is due to the AI/ML model failure.*   *Observation 13: For the container of the reported data samples in data collection, L1 signaling has lower latency and is applicable to training and monitoring, while RRC signaling has larger latency and is applicable only to training.*  *Observation 14: For the overhead of the reported data samples in data collection, L1 signaling can be comparable with L3 signaling as the monitoring window with L1 signaling can be triggered with an on-demand manner rather than always-on.*  *Proposal 18: For the potential spec impact of data collection,*   * *Both L1 signaling and RRC signaling can be considered to carry the reported data samples for model training.* * *At least L1 signaling should be considered for model monitoring to enable fast identification of AI/ML model failure.*   *Proposal 20: For the training/monitoring data collection of AI/ML model at NW side, study the methods to enable UE to feedback the RSRP labels for a subset of all measured beams in Set A to save UE report overhead.*  *Proposal 21: For the training/monitoring data collection of NW-side AI/ML model, the motivation of introducing finer resolution for UE reported RSRP measurement results for labels in Set A may be discussed after being justified in 9.2.3.1.*  *Proposal 30: Regarding the data collection for monitoring a NW-side AI/ML model at the NW, study the following options as a starting point for the contents of collected data*   * *Opt.1: UE sends M1 L1-RSRPs (corresponding to M1 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to a beam set (e.g., Set A), where M1 can be larger than 4*   + *FFS: the range of M1* * *Opt.2: UE sends M2 L1-RSRPs (corresponding to M2 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to a beam set (e.g., Set B), sends M3 L1-RSRPs (corresponding to M3 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to another beam set (e.g., Set A), where M2 and M3 can be larger than 4*   + *FFS: the range of M2, M3* * *Opt.3: UE sends M4 L1-RSRPs (corresponding to M4 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to a beam set (e.g., Set B), sends M5 beams (beam pair) based on the measurement corresponding to another beam set (e.g., Set A), where M4 can be larger than 4*   + *FFS: the range of M4, M5* |
| ZTE[4] | *Observation 15: For performance monitoring of the NW-side model, the potential specification impact of UE reporting is similar with that of the data collection.*  *Proposal 31: Model/functionality selection/activation/deactivation/switching/fallback operation should be discussed separately from performance monitoring.* |
| Vivo[5] | *Proposal 41: For BM-Case1 and BM-Case2 with a NW-side AI/ML model, regarding NW-side performance monitoring, study the following monitoring procedures:*  * UE performs resource measurement and reports corresponding measurement results including set B results and set A label data*  * NW performs beam prediction and predicted results comparison with label data to obtain performance metric(s)*  * NW makes decision(s) of model selection/activation/deactivation/switching/ fallback operation*  * Note: it can be applied on both model ID based and functionality-based LCM procedures*  *Proposal 42: Support to study hybrid-side model monitoring for BM-Case1 and BM-Case2 with a NW-side AI/ML model, which can save large measurement resources and report overhead compared to NW-side model monitoring.*  *Proposal 43: For BM-Case1 and BM-Case2 with a NW-side AI/ML model, regarding hybrid-side performance monitoring, study the following monitoring procedures:*  * UE performs resource measurement and reports set B results used for NW-side beam prediction*  * NW performs beam prediction based on set B results and indicates inference result (e.g., top-N predicted results) to UE*  * UE performs predicted result comparison with label data to obtain performance metric(s) and reports monitoring result(s) to gNB*  * NW makes decision(s) of model selection/activation/deactivation/switching/fallback operation*  * Note: it can be applied on both model ID based and functionality-based LCM procedures*  *Proposal 44: For BM-Case1 and BM-Case2 with a NW-side AI/ML model, study the potential specification impact on resource configuration for model monitoring:*  *• Specific beam pair resource configuration for Set B/Set C and/or Set A*  *• P3+P2 resource configuration that Rx beam assumption of P2 resource measurement is the best Rx beam searched from P3 procedure for performance improvement*  *Proposal 45: For BM-Case1 and BM-Case2 with a NW-side AI/ML model, study the potential specification impact on assistance information for model monitoring:*  *• Proprietary processed Rx beam information as assistance information from UE to NW, including measured Rx beam information, expected Rx beam information, and best Rx beam information.*  *Proposal 46: For BM-Case1 and BM-Case2 with a NW-side AI/ML model, study the potential specification impact on report overhead reduction for model monitoring:*  *• Reducing unnecessary L1-RSRP report where the omitted L1-RSRPs may be lower than a pre-defined threshold*  *Proposal 47: For BM-Case1 and BM-Case2 with a NW-side AI/ML model, study potential specification impact on quantization enhancement for model monitoring:*  *• High-precision L1-RSRP quantization*  *• Multi-resolution L1-RSRP quantization, e.g. high-resolution quantization for a group of best RSRPs and low-resolution quantization for others.* |
| OPPO[6] | *Observation 11: For BM-Case1 and BM-Case2 with a network-side AI/ML model and monitoring, there may be no additional specification impact on LCM.* |
| CATT[9] | *Proposal 16: Regarding the model monitoring for BM-Case1 and BM-Case2, study the specification impacts on the following aspects:*   * *Model update/switching/fallback procedures based on model monitoring results, including the signaling exchange between the gNB and UE;* * *Trigger condition for model update/switching/fallback.* |
| Ericsson[14] | *Observation 9 System/link level performance metrics based model monitoring method has low complexity and low signalling overhead. It can be sufficient for the NW to monitoring the AI-feature performance of users with MBB services if fall back operations are supported.*  *Observation 10 Monitoring of NW-sided models can be done in a step-wise approach by considering different performance metric(s) and associated performance monitoring related KPIs (e.g., latency, complexity, signaling overhead, etc.).*  *Proposal 11 Study both an RRC-message based and L1 fast CSI reporting-based data collection methods for the NW to monitor NW-sided models, including*  *• periodic UE reporting and event-triggered UE reporting configuration*  *o SSB/CSI-RS resource configuration*  *o Monitoring window configuration*  *Observation 11 Model input/output data distribution based monitoring methods should be based on data statistics of sufficient input/output data samples collected from the system.* |
| Fujitsu[15] | *Proposal 3: Considering the performance of AI/ML model and the overhead of beam reporting, study the necessity and scheme on quantization of beam reporting in the data collection, inference and monitoring on AI/ML model.* |
| Xiaomi[16] | *Proposal 19: For NW-side model monitoring for network-side AI/ML model, support to report both set B and set C, where set B will be used as network-side AI/ML model input, and set C consists of Top-K beams by UE’s measurement of set A.* |
| Samsung[19] | *Proposal 3. For BM-Case1 with a network-side AI/ML model, for model monitoring, the following aspects should be further study:*   * *UE to report the measurement results of more than 4 beams in one reporting instance* * *Assistance information (e.g., UE speed, indoor/outdoor) associated with the beam measurements*   *Proposal 13. For BM-Case1 with a UE-side AI/ML model and NW-side performance monitoring, further study the specification impacts on the following aspects:*   * *NW to provide Set A and/or Set B information to UE for measurement and/or reporting* |
| CIACT[20] | *Proposal 6: For BM-Case1 and BM-Case2 with NW-side AI/ML model monitoring, NW could configure measurement beam(pair) set and reporting periodicity to UE for AI/ML model monitoring.* |
| CMCC[22] | *Proposal 12: For BM-Case1 with a NW-side AI/ML model, study the following mechanism for model monitoring:*   * *NW-side Model monitoring*   + *NW monitors the performance metric(s)*   + *NW makes decision(s) of model selection/activation/ deactivation/switching/fallback operation* |
| MediaTek[23] | *Proposal 5: For NW-side model monitoring, the number of beams and the quantity (metric) of the report values in one reporting should be determined by the benchmark alternatives and performance metrics that are used for model monitoring.* |
| Apple[25] | *Proposal 2:*   * *For Model training at the NW side & inference at the NW side (Alt. 1), study efficient signalling of set B selection or beam selection and RSRP representation.* * *For Model training at the NW side & inference at the UE side (Alt. 3), study model generalization performance, study model transfer/model delivery for cell-specific AI models and non cell-specific AI models.* |
| Lenovo[26] | *Proposal 7: NW-side model monitoring is supported for NW-side AI/ML inference, and Rel-15 beam management procedure can be reused with necessary enhancements.* |
| DOCOMO[29] | *Proposal 9: Study the overhead reduction of L1 signalling to report SetA beam measurements for NW-based model monitoring.* |
|  |  |
|  |  |

###### DP 4.2.1

Mod’s assessment: Some proposals are quite general and the detailed proposals are quite diverging. Moreover, most of the detailed proposals are only suggested by one company. Thus, let’s wait for more inputs

|  |  |
| --- | --- |
| Company | Comments |
| vivo | We think it is needed to enable hybrid model monitoring for NW-side AI/ML model. If only NW side only monitoring is supported, UE has to report set A results which have high report overhead. Hence we think it is beneficial to let UE report some KPIs related with the inference to reduce report overhead.  Based on the discussion so far, we would like to see whether it is RAN1 understanding that it is included in NW side monitoring, which is similar as a subsequent proposal for UE side model. |
| HW/HiSi | We would like to emphasize that this discussion also is related to the data collection.  It is important that the NW can quickly find out reason for performance drop. If it is e.g. due to the AI/ML model or other reasons. It is therefore necessary to facility fast data collection (e.g. based on L1 signaling) |
| Mod | Companies are invited to share views on the proposal/suggestions from vivo and HW. |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## UE-side model

In previous RAN1 meeting(s), the agreement(s)/conclusion(s) were made as below:

|  |
| --- |
| **RAN1#110bis-e**  Agreement  For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the following alternatives for model monitoring with potential down-selection:   * Atl1. UE-side Model monitoring   + UE monitors the performance metric(s)   + UE makes decision(s) of model selection/activation/ deactivation/switching/fallback operation * Atl2. NW-side Model monitoring   + NW monitors the performance metric(s)   + NW makes decision(s) of model selection/activation/ deactivation/switching/ fallback operation * Alt3. Hybrid model monitoring   + UE monitors the performance metric(s)   + NW makes decision(s) of model selection/activation/ deactivation/switching/ fallback operation   **RAN1#112**  Agreement  For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding NW-side performance monitoring, study the following aspects as a starting point including the study of necessity:   * Configuration/Signaling from gNB to UE for measurement and/or reporting * UE reporting to NW (e.g., for the calculation of performance metric) * Indication from NW for UE to do LCM operations * Other aspect(s) is not precluded * Note1: At least the performance and reporting overhead of model monitoring mechanism should be considered   Agreement  For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding UE-side performance monitoring, study the following aspects as a starting point including the study of necessity and feasibility:   * Indication/request/report from UE to gNB for performance monitoring   + Note: The indication/request/report may be not needed in some case(s) * Configuration/Signaling from gNB to UE for performance monitoring * Other aspect(s) is not precluded |

|  |  |
| --- | --- |
| Huawei[2] | *Proposal 7: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact of AI/ML model monitoring from the following additional aspect on top of previous agreements for inference and training:*   * *Indication of the associated Set A from NW to UE, e.g., association/mapping of beams within Set A and beams within Set B if applicable.*   *Proposal 8: For the model training/monitoring/inference of the UE-side AI/ML model under BM-Case1 and BM-Case2, for how to indicate the association of beams within Set A and beams within Set B:*   * *Study the indication methods, e.g., indicating the CSI report/resource set ID, time offset, etc.* * *Study the issue when Set A has not been swept in the local cell.*   *Proposal 9: For the model training/monitoring/inference of the UE-side AI/ML model under BM-Case1 and BM-Case2, for how to indicate the mapping of beams within Set A and beams within Set B:*   * *Study the indication methods, e.g., in forms of the set of IDs, bitmap, etc.* * *Study whether/how to indicate such mapping when Set B is a set of wide beams different from Set A.*   *Observation 15: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, for Alt.1 UE-side model monitoring, it may be problematic if UE autonomously makes decisions without reporting to gNB, due to the following reasons:*   * *The UE may not be aware of all aspects impacting the AI/ML model operation.* * *NW may suffer unknown performance fluctuation.* * *gNB is not aware of the change of the model input/output if UE autonomously makes the decision of model switching/fallback, which may result in mismatched RS configurations and/or mismatched content/payload size of the expected UE report.*   *Proposal 27: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, for Alt.1 UE-side model monitoring, the UE should report the decision to the NW, and then the NW could indicate the UE a corresponding execution of the decision.*  *Proposal 28: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, for Alt.3 hybrid model monitoring, the following metrics can be studied for UE reports*   * *Per sample metric, e.g., beam prediction accuracy of each data sample.* * *Statistical metric, e.g., average, 5%-ile of the beam prediction accuracy, etc.*   *Proposal 29: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, gNB can configure a threshold criterion (e.g., in terms of threshold throughput/L1-RSRP, or threshold beam prediction accuracy) to facilitate UE to make the monitoring decision for Alt.1 (UE-side model monitoring) or make the conditional report for Alt.3 (hybrid model monitoring).* |
| H3C[3] | *Proposal 6: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, deprioritize the study of NW-side model monitoring.* |
| ZTE[4] | *Proposal 32: For UE-side model, depending on which side to calculate the performance monitoring metrics and whether the performance monitoring metrics should be reported, further study the following options:*  *• UE-side performance monitoring: performance monitoring metrics are calculated by UE, and the performance monitoring metrics are not reported to network side.*  *• Network-side performance monitoring: performance monitoring metrics are calculated by network (with/without the potential to inform UE about the performance monitoring metrics).*  *• Hybrid performance monitoring: performance monitoring metrics are calculated by UE, and then the performance monitoring metrics are reported to network side.*  *Proposal 33: Study performance monitoring mechanism on the basis of beam failure recovery mechanism in current specification.*  *Proposal 34: The final decision on model/functionality selection/activation/deactivation/switching/fallback operation should be made by NW to guarantee overall NW performance.* |
| Vivo[5] | *Proposal 48: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding Hybrid-side performance monitoring, study the following monitoring procedures:*  *• UE performs label data measurement, set B measurement and beam prediction and predicted result comparison with label data to obtain performance metric(s)*  *• NW makes decision(s) of model selection/activation/deactivation/switching/fallback operation*  *• Note: it can be applied on both model ID based and functionality-based LCM procedures*  *Proposal 49: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the potential specification impact on resource configuration for model monitoring:*  *• Specific beam pair resource configuration for Set B/Set C and/or Set A*  *• P3+P2 resource configuration that Rx beam assumption of P2 resource measurement is the best Rx beam searched from P3 procedure for performance improvement*  *Proposal 50: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the potential specification impact on assistance information for model monitoring:*  *• Proprietary processed Tx beam information as assistance information from NW to UE*  *Proposal 51: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the potential specification impact on request signaling for model monitoring:*  *• Resource request signaling for data collection from UE to NW*  *- Beam pair resources request for model monitoring purpose including the number of requested labels, and potentially some associated triggering events to be defined*  *- P3+P2 beam sweeping resources request for model monitoring purpose including the number of requested labels, and potentially some associated triggering events to be defined*  *• Minimum resource number request for data collection from UE to NW*  *- Minimum number of requested beams for model monitoring w or w/o resource request signaling*  *- Minimum number of requested repetitions for model monitoring w or w/o resource request signaling*  *Proposal 52: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the potential specification impact on monitoring report for model monitoring:*  *• Monitoring result report from UE to NW, including label data report or performance metric report* |
| OPPO[6] | *Proposal 14: For BM-Case1 and BM-Case2 with UE-side model, study the (Alt1) UE-side model monitoring as a starting point.* |
| Nokia[8] | *Proposal 6. For UE-sided BM-Case1, for any functionality configured towards the UE, the gNB shall be able to consider the performance monitoring at the NW side.*  *• A dedicated beam measurement and reporting configuration that enables measurement and reporting of full/partial Set A (associated with a given functionality) can be used to enable performance monitoring at the NW side.*  *Proposal 7. For UE-sided BM-Case1, for any functionality activated towards the UE, the gNB shall be able to configure the performance monitoring at the UE side.*  *• A dedicated beam measurement and reporting configuration that enables measurement and reporting of full/partial Set A (associated with a given functionality) can be used to enable performance monitoring at the UE side.*  *• The UE may consider a performance monitoring KPI (Top-K/1 beam accuracy) with gNB configured threshold to determine functionality failures of the activated functionality.*  *o further study the framework of functionality failures detection for an activated functionality*  *o further study the reporting framework for functionality failures.*  *Proposal 14. For UE-sided BM-Case2, for any functionality configured towards the UE, the gNB shall be able to consider the performance monitoring at the NW side.*  *• A dedicated beam measurement and reporting configuration that enables measurement and reporting of full/partial Set A (associated with a given functionality) can enable performance monitoring at the NW side.*  *Proposal 15. For UE-sided BM-Case2, for any functionality activated towards the UE, the gNB shall be able to configure the performance monitoring at the UE side.*  *• A dedicated beam measurement and reporting configuration that enables measurement and reporting of full/partial Set A (associated with a given functionality) can be used to enable performance monitoring at the UE side.*  *• The UE may consider a performance monitoring KPI (Top-K/1 beam accuracy) with gNB configured threshold to determine functionality failures of the activated functionality.*  *o further study the framework of functionality failures detection for an activated functionality*  *o further study the reporting framework for functionality failures.* |
| CATT[9] | *Proposal 14: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, for Alt.1 UE-side model monitoring, study the potential specification impacts on the following aspects:*   * *Reporting the decision of model activation/ deactivation/switching/fallback to the network;* * *Acknowledgement mechanism of model activation/ deactivation/switching/fallback from the network.* |
| Intel[10] | *Proposal 2: For data collection and UE-side model performance monitoring with AI/ML model at UE side, support UE triggered reference signal transmission from the gNB to enable the UE to perform L1 measurements at least on Set B for both BM-Case 1 and 2.*  *Proposal 6: For UE-side AI/ML model, support UE-side and Hybrid model monitoring. NW-side monitoring can be further studied and used in specific cases if the model is transferred from the NW to the UE.* |
| IDC[11] | *Observation 10: Definition of ‘monitoring’ in the agreement for model monitoring is not clear enough.*  *Proposal 13: Clarify the details of ‘monitoring for each alternative including UE reporting of the performance metric(s) for Alt3.*  *Proposal 16: For configuration/signaling from gNB to UE, consider configuration of monitoring RS/channel, evaluation methodology for monitoring and confirmation on UE request/trigger.*  *Proposal 17: For indication/request/report from UE to gNB, consider reporting UE monitoring result and trigger of a model recovery procedure.* |
| Ericsson[14] | *Proposal 12 For BM-Case1 and BM-Case2 with a UE-side AI/ML model, conclude that UE is responsible to make sure the model is working correctly.*  *Proposal 13 Study mechanisms for the NW to configure a UE operating an AI/ML based beam prediction model to fallback to a legacy non-ML based beam reporting.* |
| Fujitsu[15] | *Proposal 10: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, the Alt.2 and Alt.3 are suggested to be prioritized for further study of model monitoring.*  *Proposal 11: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the potential specification impacts of hybrid model monitoring on*   * *The configuration of filtering on the performance metric calculation* |
| Xiaomi[16] | *Proposal 14: For UE-side AI/ML model with UE-side model monitoring, support UE to indicate the decision to NW.*  *Proposal 15: For UE-side AI/ML model with NW-side model monitoring, consider UE to report the beam (pair) ID of predicted Top-K beams and genie-aided Top K beams, and/or genie-aided L1-RSRP and predicted L1-RSRP of predicted Top-K beams to NW for performance metric calculation.*  *Proposal 16: For UE-side AI/ML model with hybrid model monitoring, support an event-triggered report of performance metric from UE based on a threshold configured by gNB.*  *Proposal 18: For UE-side AI/ML model, UE-side initiated performance monitoring based on RA or SR can be considered, and NW-side initiated performance monitoring based on measurement configuration via RRC can be considered.* |
| LGE[18] | *Proposal #10: For UE-sided AI/ML model, Alt1(UE-side model monitoring) should be supported.*  *Proposal #11: Further study whether dedicated signaling or procedure for UE-side performance monitoring is necessary by considering that Set A beams could be provided based on UE capability report.*  *Proposal #12: For NW-sided model monitoring, UE reporting based on Set A can be considered but with an assumption that measurement and reporting on Set A or potential beams of Set A shall happen rarely.* |
| Samsung[19] | *Proposal 12. For BM-Case1 with a UE-side AI/ML model, Alt2 (i.e., NW-side model monitoring) and Alt3 (i.e., Hybrid model monitoring) are preferred.* |
| CIACT[20] | *Proposal 4: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding UE-side performance monitoring, the periodicity of UE measurements on L1-RSRP of predicting beam pairs and/or all beam pairs in Set A could be achieved from NW.*  *Proposal 5: For UE-side AI/ML model monitoring, UE side directly monitoring (Alt.1) should be baseline.* |
| CMCC[22] | *Proposal 10: For BM-Case1 with a UE-side AI/ML model, study the decision report and acknowledgement mechanism for UE-side model monitoring.*  *Proposal 11: For BM-Case1 with a UE-side AI/ML model, study the information needed for UE reporting to NW to calculate the performance metric for NW-side model monitoring.* |
| MediaTek[23] | *Proposal 6: For NW-side performance monitoring for UE side model, focus on the discussion when the performance metric is Alt.2 to Alt.4.*  *Proposal 7: To avoid large UE reporting overhead, Alt.1 of the benchmark should be deprioritized for NW-side performance monitoring for UE side model.*  *Proposal 8: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding Alt.3 (Hybrid model monitoring), we support proposal 5.3.3 with the following updates:*  *For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding Alt.3 (Hybrid model monitoring), study the following aspects as a starting point including the study of necessity:*   * *Signaling from gNB to UE for performance monitoring (e.g., dedicated RS configuration for measurement)* * *The contents of UE reporting and the UE reporting mechanism to NW* * *The NW-side control level of the model selection/activation/deactivation/switching/fallback* * *Other aspect(s) is not precluded*   *Note: At least the performance and reporting overhead of model monitoring mechanism should be considered* |
| Lenovo[26] | *Proposal 8: For both NW and UE-side performance monitoring for UE-side AI/ML inference, at least support aperiodic beam measurement for model monitoring and dynamic beam change within the beam set associated with beam measurement should be further studied.* |
| Qualcomm[27] | *Proposal 7*  *For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding UE-side performance monitoring, study the following signalling aspects related to configuration/signalling from gNB to UE for performance monitoring:*  *• Dedicated RS from gNB to UE for performance monitoring* |
| NEC[28] | *Proposal 6:* *For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding Alt.3 (Hybrid model monitoring), study the following information of UE reporting and corresponding reporting mechanism.*   * *Performance metric.* * *Non-performance metric, which is determined based on the monitored performance metric(s).* |
| DOCOMO[29] | *Proposal 11: Study the hybrid performance monitoring, where NW obtains the performance matric calculated at UE.* |
|  |  |

###### Proposal 4.3.1

In RAN1#110bis-e, three alternatives were agreed as the candidate solutions for the monitoring of UE-side AI/ML model:

* Alt.1: UE-side model monitoring
* Alt.2: NW-side model monitoring
* Alt.3: Hybrid model monitoring

In RAN1#112, further refinements based on the above agreement were made and two agreements were achieved for both model-ID-based and functionality-based LCM:

* UE-side performance monitoring
* NW-side performance monitoring

Based on the tdocs, many companies support Alt.3 of RAN1#110bis-e. Meanwhile, some companies think the new agreements of RAN1#112 is not clear and suggest to clarify Alt.3 of RAN1#110bis-e is included in “NW-side performance monitoring”. Thus, Proposal 4.3.1 is provided to make the clarification.

The related proposals in tdocs are as below:

* ZTE: Proposal 32
* Vivo: Proposal 48
* Intel: Proposal 6
* IDC: Proposal 17
* Fujitsu: Proposal 10, 11
* Xiaomi: Proposal 16
* Samsung: Proposal 12,
* MediaTek: Proposal 8
* NEC: Proposal 6
* DCM: Proposal 11

***Proposal 4.3.1:******For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding hybrid performance monitoring, study the following aspects as a starting point including the study of necessity:***

* ***Configuration/Signaling from gNB to UE for measurement and/or reporting***
* ***UE calculates the performance metric(s) and report it to NW***
* ***Indication from NW for UE to do LCM operations***
* ***Other aspect(s) is not precluded***
* ***Note1: At least UE complexity and power consumption, performance, ~~and~~ reporting overhead, and latency of model monitoring mechanism should be considered***

|  |  |
| --- | --- |
| Company | Comments |
| Nokia/NSB | This is OK. Partly covered in the last meeting agreement.  Agreement  For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding NW-side performance monitoring, study the following aspects as a starting point including the study of necessity:   * Configuration/Signaling from gNB to UE for measurement and/or reporting * UE reporting to NW (e.g., for the calculation of performance metric) * Indication from NW for UE to do LCM operations * Other aspect(s) is not precluded * Note1: At least the performance and reporting overhead of model monitoring mechanism should be considered   Mod: Some companies think for the highlighted part, the only example is associated with “Atl2. NW-side Model monitoring” and it is not clear whether hybrid mechanism is included or not. The intention of this proposal is to clarify the “UE reporting to NW” include both the previous agreements “Atl2. NW-side Model monitoring” and “ Alt3. Hybrid model monitoring” |
| Google | We do not think the UE needs to report the performance metric. It can be actually a simple ACK/NACK report for the TCI indication. Now we have a ACK/NACK report to indicate whether UE receives the TCI indication, but we do not have a ACK/NACK to report whether the indicated TCI is right.  Mod: That is one design. Companies proposed some other designs |
| New H3C | OK |
| NTT DOCOMO | Support the proposal. |
| LG | The proposed case is ambiguous with respect to what UE needs to report. If it is already covered by previous agreement as Nokia pointed out, we don’t see necessity of having this additional agreement.  Mod: Please see the reply to Nokia and also vivo’s comment. |
| vivo | We are fine with the proposal. We think it is needed to have such proposal. The previous one seems not clear whether UE reporting of performance metrics is included. It is needed to align RAN1’s understanding. |
| HW/HiSi | [ok], could it please be explained what is the difference to the hybrid mode here?  Mod: Please see the reply to Nokia and also vivo’s comment. The agreements in the last meeting includes more details and “model monitoring” is changed to “performance monitoring” to avoid the potential misleading that it is only applicable for model-ID-based LCM |
| Spreadtrum | Support the proposal. |
| CAICT | Support the proposal. |
| CATT | OK |
| CMCC | In our understanding, regarding NW-side performance monitoring, UE report measurement results and NW calculates performance metric. If UE calculates the performance metric(s) and reports it to NW, this is more like hybrid model monitoring.  Mod: Please see the reply to Nokia/Huawei and also vivo’s comment. |
| Ericsson | OK |
| Xiaomi | We can support this proposal. but it is better to clarify that in this case, what is the difference between NW-side model monitoring and hybrid model monitoring?  Mod: Please see the reply to Nokia/Huawei and also vivo’s comment. |
| Samsung | As mentioned by Nokia, this is somehow agreed in the last meeting. Not sure the necessity of this proposal.  Mod: Please see the reply to Nokia/Huawei and also vivo’s comment. |
| Sony | Support |
| Fujitsu | Could FL clarify the boundary of NW-side performance monitoring and hybrid performance monitoring? From our understanding, the performance metrics reported by UE is more like the scope of hybrid monitoring.  Mod: Please see the reply to Nokia/Huawei and also vivo’s comment. |
| NVIDIA | Support. |
| Qualcomm | As mentioned by Nokia, this is essentially covered in the cited previous agreement.  Mod: Please see the reply to Nokia/Huawei and also vivo’s comment. |
| Lenovo | Support |
| MediaTek | If UE calculates the performance metrics, then what is the difference between NW-side performance monitoring and Hybrid performance monitoring for UE-side model?  Mod: Please see the reply to Nokia/Huawei and also vivo’s comment. |
| ZTE | We are also confusing about the boundary of NW-side performance monitoring and hybrid performance monitoring. Note that in previous agreement, the definition of NW-side performance monitoring is that NW monitors the performance metric(s).  Mod: Please see the reply to Nokia/Huawei and also vivo’s comment. |
| NEC | OK |
| MediaTek | We understand the intension now. We think the wording can be more specific here to avoid any confusion. Suggest the following change:  ***Proposal 4.3.1:******For BM-Case1 and BM-Case2 with a UE-side AI/ML model, the aspect regarding UE reporting to NW for NW-side performance monitoring agreed in RAN1#112 also includes the following example:***   * ***UE calculates the performance metric(s) and reports it to NW*** |
| InterDigital | Fine with the proposal. |
| Futurewei | Support |
| Qualcomm | Do not support. It is not really clear why “Alt3. Hybrid model monitoring” should be categorized under the umbrella of “NW-side performance monitoring”. Let’s review the agreement from RAN1 #112:  Agreement  For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding UE-side performance monitoring, study the following aspects as a starting point including the study of necessity and feasibility:   * Indication/request/report from UE to gNB for performance monitoring   + Note: The indication/request/report may be not needed in some case(s) * Configuration/Signaling from gNB to UE for performance monitoring * Other aspect(s) is not precluded   As we see in the agreement, UE report to gNB for performance monitoring is also included under the umbrella of UE-side performance monitoring. With the current wording of the proposals, there is no crisp and clear boundary between UE-side and NW-side performance monitoring, and no easy way to categorize the so-called “hybrid monitoring” strictly under the umbrella of the two agreed flavors for performance monitoring. If *UE* is monitoring/computing the KPIs and reporting them to NW, it is not clear why it should be called *NW*-side monitoring. |
| LG | Similar view as QC. If UE needs to calculate performance metric, why we categorize it as NW-side monitoring? The monitoring itself is performed by UE, although the final decision (e.g. enabling/disabling functionality) is performed by NW. |
| Mod | Let’s try another way. A new proposal is suggested to explicitly mention hybrid performance monitoring”, which is based on the agreement about “NW-side performance monitoring”. In my understanding, the meaning of the new proposal is the same as the original one. The changes are highlighted by RED. |
| Nokia/NSB | To our reading, there is no hybrid monitoring anymore. There are only two-modes, NW monitor the performance (with or without UE reporting metrics) and UE monitor performance to deactivate a functionality (or background model switching as UE like).  ***Proposal 4.3.1:******For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding NW-side ~~hybrid~~ performance monitoring, study the following aspects as a starting point including the study of necessity:***   * ***Configuration/Signaling from gNB to UE for measurement and/or reporting*** * ***UE calculates the performance metric and report it ~~reporting~~ to NW ~~(e.g., for the calculation of performance metric)~~*** * ***Indication from NW for UE to do LCM operations*** * ***Other aspect(s) is not precluded*** * ***Note1: At least the performance and reporting overhead of model monitoring mechanism should be considered*** |
| Mod | @Nokia: As you can see from the comments, most companies think there is a hybrid scheme, which is only a small different between NW-side monitoring. No matter it is merged in to NW-side monitoring or has a separate name, the scheme is always there. So does the potential spec impact. Could you live with the proposal? |
| ZTE | Fine with the proposal. Then, the difference between hybrid and NW-side performance monitoring is which side (i.e., UE or NW) is responsible for the calculation of the performance metric, right?  Mod: Yes |
| Xiaomi | As for the reporting of the performance metric, we suggest to add an FFS: periodical report or event-triggered report.  Mod: The details can be discussed later |
| HW/HiSi | Support |
| Qualcomm | We have a comment about the note. Because such kind of monitoring could be event-triggered (which is not clear from the current wording), we suggest adding the following to the note:  ***Proposal 4.3.1:******For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding ~~NW-side~~ hybrid performance monitoring, study the following aspects as a starting point including the study of necessity:***   * ***Configuration/Signaling from gNB to UE for measurement and/or reporting*** * ***UE calculates the performance metric and report it ~~reporting~~ to NW ~~(e.g., for the calculation of performance metric)~~*** * ***Indication from NW for UE to do LCM operations*** * ***Other aspect(s) is not precluded***   ***Note1: At least the UE complexity and power consumption, performance, ~~and~~ reporting overhead, and latency of model monitoring mechanism should be considered*** |
| LG | Ok to call this as ‘hybrid performance monitoring’ although we haven’t define it for performance monitoring (note that previous agreement for categorizing monitoring was for model monitoring)  However, the current wording seems to cover only Case 1 below, and not cover Case2.  **Case1)**   1. UE measures performance metric(s) and report performance metric(s) (e.g. RSRP, BLER, SINR, etc) periodically to NW 2. NW makes decision on functionality selection/activation/ deactivation/switching/ fallback operation   **Case2)**   1. UE measures performance metric(s) and determine whether there is an event. When the event happens, UE reports NW that the occurrence of the event. 2. NW makes decision on functionality selection/activation/ deactivation/switching/ fallback operation   To cover both cases, the second bullet needs to be revised, e.g.   * ***UE calculates the performance metric and report the performance metric or an event related to the performance metric to NW***   Mod: In my understanding, Case2 belongs to UE-side performance monitoring (please see Mod’s notes in 4.3.2, where companies are encouraged to share views). Let’s hear more views. |
| NEC | Support in general. For second bullet, we think that UE may calculate one or multiple performance metrics and report it to NW. Therefore, we prefer to add “(s)” as follows:   * ***UE calculates the performance metric(s) and report it to NW***   Mod: Updated |
| Samsung3 | Ok with the latest update. |
| Ericsson | Support latest update. |
| Futurewei | Support |
| ZTE | Support latest update. |
| LG | Based on the agreed 3 categories for model monitoring, we think that it is better to include Case 2 in our previous response within the category of ‘hybrid performance monitoring’. Suggest the wording update as proposed previously. |
| CMCC | Ok. |

###### Proposal 4.3.2(Closed)

Companies have different preference on the alternatives. Based on the submitted tdocs, companies’ views are collected in the following table:

|  |  |  |
| --- | --- | --- |
|  | Support/Prefer | Deprioritized |
| NW-side performance monitoring | Huawei, ZTE, vivo, Nokia, CATT, Intel (support hybrid), Fujitsu, Xiaomi, Samsung, DCM, Spreadtrum, Ericsson | H3C, |
| UE-side performance monitoring | H3C, vivo, OPPO, CATT, Intel, Ericsson, Xiaomi, LGE, CIACT, QC, Spreadtrum |  |

There are two LCM framework discussed for AI/ML operations:

* Model-ID-based LCM
* Functionality-based LCM

For different LCM frameworks, the preferred performance monitoring mechanism may be different. Thus, Proposal 4.3.2 and 4.3.3 are suggested for functionality-based LCM and model-ID-based LCM, respectively.

***Proposal 4.3.2:******For BM-Case1 and BM-Case2 with a UE-side AI/ML model, when functionality-based LCM is applicable, support NW-side performance monitoring to facilitate functionality related LCM operations (i.e., activation/deactivation/fallback/switching of AI/ML functionality)***

* ***FFS: UE-side performance monitoring***

|  |  |
| --- | --- |
| Company | Comments |
| Nokia/NSB | Support |
| Google | Do not support. UE-side model should be monitored by the UE.  Mod: This proposal is for the functionality, rather than for AI model. Legacy functionalities of NR are usually controlled by NW. Thus, most companies think the similar principle is also applicable here. |
| New H3C | We open to discuss about this. But we need point out that if NW monitors the performance, UE need to report the L1-RSRP of Set A beams to NW to calculate the performance, which will bring additional reporting overhead.  Mod: Agree |
| NTT DOCOMO | The clarification is needed about what performance is monitored (e.g., performance of individual model? Or performance of functionality?). If it is about the performance of functionality, NW-side performance monitoring should be supported. On the other hand, the performance of models within a functionality should be monitored at UE side transparently to NW side, in case of functionality-based LCM.  Mod: The proposal is updated. Please check whether it is ok |
| LG | For functionality-based LCM, likewise current specification, it is NW to enable/disable functionality, but it is UE’s responsibility to meet certain performance requirement. In this sense, we support UE-side performance monitoring only.  Mod: sorry. Not get the point. Yes, agree “it is UE’s responsibility to meet certain performance requirement”, but failed to see how it is related to “UE-side performance monitoring” |
| vivo | We haven’t agreed on the support of functionality based LCM. Hence we suggest the following change.  ***Proposal 4.3.2:******For BM-Case1 and BM-Case2 with a UE-side AI/ML model, when functionality-based LCM is*** ***agreed to be feasible and applicable, support NW-side performance monitoring***  Mod: If it is not feasible, it will not applicable.  For UE side monitoring of functionality based LCM, although most of the components can be implementation, but at least the RS configuration from NW should be specified. Is this common understanding? If so, we suggest add a new bullet.  ***For BM-Case1 and BM-Case2 with a UE-side AI/ML model, when functionality-based LCM is*** ***agreed to be feasible and applicable, support signaling exchange on RS configuration for UE-side performance monitoring.***  Mod: If moderator remembered correctly, the new part is included in previous agreement |
| HW/HiSi | Seems Ok on first sight |
| Spreadtrum | UE-side monitoring can also be considered when functionality-based LCM is applicable.  Mod: FFS part is added |
| CAICT | Open to discuss and the relationship between UE side monitoring and NW-side monitoring needs further discussions. |
| CATT | Support |
| CMCC | If NW-side performance monitoring applies, we wonder how to perform model switching since functionality-based LCM only allows NW to perform function level action.  Mod: This proposal is only for functionality performance monitoring. The issue regarding model switching under a functionality is a separate discussion. There are some related discussions in AI 9.2.1 |
| Ericsson | Think it is ok. |
| Xiaomi | We also support UE-side performance monitoring.  Mod: FFS part is added |
| Samsung | Support |
| Sony | the proposal is OK when there is only one model with each function/feature at UE-side. But we have concerns that when functionality-based LCM is applicable, if multi models correspond to one function/feature, NW can’t make the decision based on the performance of UE-side single model.  Mod: It is another issue whether/how to do model-level operations within a functionality. There are some discussions in AI 9.2.1 |
| NVIDIA | Support. |
| Qualcomm | Do not support. This should not be a binary determination. For functionality-based LCM, we can have both UE-side and network-side performance monitoring. UE may monitor the performance of AI/ML models for a given functionality and may switch among multiple AI/ML models within the same functionality, and network may monitor the performance of UE “at a functionality level”.  Mod: This proposal is only for functionality performance monitoring. The issue regarding model switching under a functionality is a separate discussion. There are some related discussions in AI 9.2.1 |
| Lenovo | We think it’s better to discuss this proposal after the performance metrics is agreed.  Mod: Would you like to elaborate a bit more what’s the issue/concern here? |
| Apple | The UE should be able to perform performance monitoring autonomously, and NW is not involved. Don’t support this proposal.  Mod: This proposal is only for functionality performance monitoring. Legacy functionalities of NR are usually controlled by NW. Thus, most companies think the similar principle is also applicable here. Would you like to elaborate a bit more on your concern? |
| MediaTek | In our opinion UE may have the preferred functionality conditions based on the observed performance metrics or UE’s current status. For example, depending on UE speed, the UE can activate/deactivate different number of future instances or measurement periodicity for BM Case2. Different functionality conditions may have different preferred monitoring mechanisms.  Mod: That is one possibility. By reading the tdocs, it seems most companies still prefer the same principle that is used for legacy functionality. Let’s hear more views. An FFS part is also added |
| ZTE | Support. |
| NEC | Support |
| HW/HiSi | To my understanding, the support mentioned by previous companies was for U1 (without the FFS for UE side monitoring).  For my understanding, based on what discussion has the FFS been added in U2?  U1:  ***Proposal 4.3.2:******For BM-Case1 and BM-Case2 with a UE-side AI/ML model, when functionality-based LCM is applicable, support NW-side performance monitoring to facilitate functionality related LCM operations***  Mod: As you can see, several companies suggest UE performance monitoring |
| MediaTek | We are fine with putting an FFS for UE-side performance monitoring, a note maybe better. Maybe the following wording is better?   * ***Note: The applicable UE-side performance monitoring cases under*** ***functionality-based LCM is a separate discussion*** |
| InterDigital | In our view, it is premature to discuss this proposal. It’s better to discuss after deciding which type of LCM is supported. |
| Futurewei | OK to us. |
| Qualcomm | Do not support. For functionality-based LCM, UE-side performance monitoring and NW-side performance monitoring can happen in coexistence with each other, and UE-side performance monitoring does NOT preclude NW-side performance monitoring, and vice versa. Let’s look at the following agreement from RAN1 #112:  Agreement  For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding UE-side performance monitoring, study the following aspects as a starting point including the study of necessity and feasibility:   * Indication/request/report from UE to gNB for performance monitoring   + Note: The indication/request/report may be not needed in some case(s) * Configuration/Signaling from gNB to UE for performance monitoring * Other aspect(s) is not precluded   There may be two UE-side AI/ML models for a given beam prediction functionality. The AI/ML models may be transparent to NW and NW knows UE supports this functionality. To monitor the performance of any of these two AI/ML models within that functionality, UE may request signaling such as dedicated RS from NW and can monitor the performance. The UE may decide to switch among those two AI/ML models, e.g., based on the result of UE-side monitoring. This switching may be transparent to NW.  Mod: It is related to model operations. Not related to functionality operations. Thus, what’s the connection between it and Proposal 5.3.2  For this scenario, even if the NW may not know about the AI/ML models or switching among them, NW may still activate or deactivate the AI/ML functionality (NW-side performance monitoring), at least based on some of the KPIs agreed for monitoring in the previous meetings (e.g., indirect KPIs).  Mod: As you said, it is NW-side performance monitoring  This illustrative example is trying to highlight the fact that we cannot exclusively associate NW-side or UE-side monitoring to functionality-based LCM. Both flavors of monitoring could be applicable.  Mod: The proposal is just related to functionality-related LCM operations. The “i.e., activation/deactivation/fallback/switching of AI/ML functionality” is added to emphasize it. |
| ZTE | We support NW-side performance monitoring to facilitate functionality related LCM operations. The execution of the final decision about functionality operations would normally affect the configuration of CSI resource and reporting, which should definitely be NW controllable. |
| CATT | Support.  In AI 9.2.1, we have the following agreements  Agreement  For UE-side models and UE-part of two-sided models:  …   * In functionality-based LCM   + Network indicates activation/deactivation/fallback/switching of AI/ML functionality via 3GPP signaling (e.g., RRC, MAC-CE, DCI).   + Models may not be identified at the Network, and UE may perform model-level LCM.     - Study whether and how much awareness/interaction NW should have about model-level LCM   ….  It means the indication of functionality control is always by NW, hence the NW-side performance monitoring is necessary. |
| Lenovo | According to the agreement in AI 9.2.1(thanks CATT), we support the updated proposal. |
| Xiaomi | We are fine with the updated proposal |
| vivo | We share similar view as QC. Our understanding on the current proposal is to identify what can be used and the necessary specification impact for functionality based LCM. Both UE side monitoring and NW side monitoring are applicable.  For NW side monitoring, the necessary specification impact includes RS configuration and relevant UE metric reporting.  For UE side monitoring, at least the necessary RS request from UE and RS configuration from gNB are needed for specification impact. We can further discuss whether certain KPI report from UE for functionality switch is needed.  Therefore, we suggest to revise the proposal as  ***Proposal 4.3.2:******For BM-Case1 and BM-Case2 with a UE-side AI/ML model, when functionality-based LCM is applicable, support to study both NW-side performance monitoring and UE-side performance monitoring to facilitate functionality related LCM operations***   * ***FFS: ~~UE-side performance monitoring~~ necessary specification impact for NW side and UE side monitoring (e.g., RS request/configuration, UE reporting, etc.)***   Mod: Please see the reply to QC |
| Sony | This proposal is OK for us, when functionality-based LCM is applicable, for UE-side model, it seems more reasonable that UE make the decision, but we also think it depends on the granularity of functionality and other related information, so we support to further study this proposal. |
| Fujitsu | With the agreement of AI9.2.1, we are fine to this proposal. |
| LG | Not support. As commented earlier, monitoring itself should be done by UE although functionality enabling/disabling is done by NW. For example, radio link monitoring (RLM) is done by UE in current spec  Mod: If we can agree Proposal 4.3.1, hybrid performance monitoring can be added. In that case, your concern seems be addressed. |
| Mod | The part “(i.e., activation/deactivation/fallback/switching of AI/ML functionality)” is added, the same wording as 9.2.1 agreement (please see CATT’s comment)  @all  If we can agree Proposal 4.3.1, hybrid performance monitoring can be added. That means support both hybrid and NW-side performance monitoring.  @Proponents of UE-side performance monitoring  If hybrid performance monitoring is added, would you still request to support UE side performance monitoring? |
| CMCC | Fine with network side monitoring for functionality-based LCM. |
| Nokia/NSB | Support. Hybrid monitoring mode is within the NW-sided monitoring. |
| LG | Not support. The agreement captured by CATT is about functionality activation/deactivation/fallback/switching, it is NOT about functionality monitoring shall be done by NW or UE.  In the last meeting, we have the following agreement:  Agreement  For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding NW-side performance monitoring, study the following aspects as a starting point including the study of necessity:   * Configuration/Signaling from gNB to UE for measurement and/or reporting * UE reporting to NW (e.g., for the calculation of performance metric) * Indication from NW for UE to do LCM operations * Other aspect(s) is not precluded * Note1: At least the performance and reporting overhead of model monitoring mechanism should be considered   Do we have any agreement on the above aspect? No!  If we align understanding on what is additionally required for NW-side monitoring compared to current beam operation, e.g. what kind of measurement/reporting is needed and how frequently for NW-side performance monitoring, our position may change later but we cannot agree on the current proposal without knowing the consequence of it. It may increase UE measurement/reporting overhead/complexity quite significantly (e.g. when NW wants to monitor UE performance very frequently) which is very risky to agree on this at this moment.  Mod: If we can agree Proposal 4.3.1, hybrid performance monitoring can be added. In that case, your concern seems be addressed. Is that correct understanding? |
| LG2 | Just to reply to FL:  My answer is no. It seems that the understanding of ‘hybrid’ performance monitoring is not aligned.  Let’s assume the following two cases:  **Case1)**   1. UE measures performance metric(s) and report performance metric(s) (e.g. RSRP, BLER, SINR, etc) periodically to NW 2. NW makes decision on functionality selection/activation/ deactivation/switching/ fallback operation   **Case2)**   1. UE measures performance metric(s) and determine whether there is an event. When the event happens, UE reports NW that the occurrence of the event. 2. NW makes decision on functionality selection/activation/ deactivation/switching/ fallback operation   Our understanding is that Case1 belongs to NW-side monitoring since it is NW to monitor UE performance (i.e. whether there is any problem or not is determined in NW), and Case2 belongs to UE-side monitoring since it is UE to monitor UE performance (i.e. whether there is any problem or not is determined in UE). In Case2, NW does not monitor UE performance, it only be reported by UE only when a pre-defined event happens to the UE.  With above understanding, we don’t support Case1 NW-side/hybrid monitoring as it requires UE to keep measuring/reporting something incurring large overhead. |
| NTT DOCOMO | We feel the discussion gets unnecessarily complicated.  In functionality-based LCM, the monitoring of model level is performed by UE, while the monitoring of functionality level is performed by NW based on 9.2.1. Maybe the current proposal is controversial because companies have different understanding on what to be monitored in the proposal.  ***Proposal 4.3.2:******For BM-Case1 and BM-Case2 with a UE-side AI/ML model, when functionality-based LCM is applicable, support NW-side performance monitoring at the functionality scale (i.e., activation/deactivation/fallback/switching of AI/ML functionality) and support UE-side performance monitoring at the model scale (i.e., activation/deactivation/switching of model within a functionality). ~~to facilitate functionality related LCM operations (i.e., activation/deactivation/fallback/switching of AI/ML functionality)~~***   * ***~~FFS: UE-side performance monitoring~~***   Also, we think the moderator uses hybrid performance monitoring in different manner as the past agreement.  ・hybrid performance monitoring definition A: metric is calculated at UE, and the upcoming operation is determined at NW  ・hybrid performance monitoring definition B: model level performance monitoring is performed at UE, and functionality level performance monitoring is performed at NW.  We think definition A is aligned with the agreement, while we observe the definition B is used in this proposal. |
| CATT | We support FL proposal.  According to the agreement in 9.2.1, for functionality-based LCM for UE-side models, UE **may** perform model-level LCM. In our mind, it depends on UE whether model-level UE-side performance monitoring is needed. For example, if there is only one model at UE side, the model-level performance monitoring at UE may not need. We prefer to keep UE-side performance monitoring as FFS. |
| Xiaomi | We are fine with the updated proposal |
| Nokia/NSB | Support the proposal.  To address some concerns raised above, please see our understanding.   * Functionality monitoring can happen at both UE and NW sides. * For NW sided monitoring, we consider that UE deriving metrics and reporting is included as well (i.e., hybrid mode in older agreements). * For the UE sided monitoring, the UE is allowed to monitor the performance of a functionality and it is a separate discussion. For UE-side monitoring, we shall not consider that mode as a hybrid mode as performance monitoring is mainly for the model LCM within a functionality (in many cases, this operation is transparent to the NW). Anyways, the UE can monitor the functionality to deactivate the functionality (this shall be indicated to the gNB) or to help with other operations of LCM. |
| MediaTek | Regarding hybrid performance monitoring, we agree with DCM that hybrid performance monitoring definition A aligns with the agreements. To avoid any misunderstanding, we should avoid use hybrid performance monitoring definition B during the discussion.  We think that model switch can be performed at UE and functionality level performance monitoring is performed at NW, in functionality-based LCM. Therefore, this proposal does not preclude UE to switch model based on its preference. We support the current shape of the proposal. |
| Qualcomm | Share similar views as LG2 and LG above. |
| Futurewei | Support |
| Mod | All companies are ok to NW-side and hybrid performance monitoring.  To check companies’ view:  Regarding UE-side performance monitoring for functionality related LCM operations, is the following is acceptable to companies  ***support UE-side performance monitoring***   * ***UE determine whether an event happens or not*** * ***UE reports the occurrence of the event to NW*** * ***NW makes decision on functionality related LCM operations (i.e., activation/deactivation/fallback/switching of AI/ML functionality)*** * ***FFS: how to define the event***   The “event” maybe something like beam failure (e.g., the performance is not good enough). We can discuss it later. |
| CATT | We support the performance monitoring mechanism listed by the FL. But we think it should be hybrid performance monitoring. Based on the discussion, we are totally confused about what is UE-side performance monitoring. According to the agreement in RAN1#110bis, for UE-side model monitoring, UE make decision(s) of model selection/activation/ deactivation/switching/fallback operation. We think this is also applied for UE-side performance monitoring. But in LG and FL’s example, for UE-side performance monitoring, NW makes decision on functionality related LCM operations.  ***Agreement***  ***For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the following alternatives for model monitoring with potential down-selection:***   * ***Atl1. UE-side Model monitoring***   + ***UE monitors the performance metric(s)***   + ***UE makes decision(s) of model selection/activation/ deactivation/switching/fallback operation*** * ***Atl2. NW-side Model monitoring***   + ***NW monitors the performance metric(s)***   + ***NW makes decision(s) of model selection/activation/ deactivation/switching/ fallback operation*** * ***Alt3. Hybrid model monitoring***   + ***UE monitors the performance metric(s)***   ***NW makes decision(s) of model selection/activation/ deactivation/switching/ fallback operation*** |
| vivo | It seems companies may have different view on what UE-side, NW-side monitoring or hybrid monitoring is. Then maybe we can skip these terms and just list the technical components to be studied? For example,   * RS request/configurations for monitoring * Potential KPI reporting from UE to network * NW or UE make decisions on model or functionality selection/activation/ deactivation/switching/ fallback operation |

###### Proposal 4.3.3 (Closed)

As we can see from the above table that summarizes companies’ view, both alternatives are supported by a group of companies. One compromised way may be to consider both alternatives.

***Proposal 4.3.3:******For BM-Case1 and BM-Case2 with a UE-side AI/ML model, when model-ID-based LCM is applicable, support the following alternatives for model monitoring with potential further down-selection:***

* ***UE-side performance monitoring***
* ***NW-side performance monitoring***

|  |  |
| --- | --- |
| Company | Comments |
| Nokia/NSB | There is no agreement yet to support 3GPP signalling based model-ID-based LCM. We have an agreement that model-LCM handled by the UE.  Mod: “***is applicable***” is included in the main bullet. If AI 9.2.1 will have an agreement not to support it, then it is not applicable.  We have the following suggestion to align with 9.2.1 discussion,  ***Proposal 4.3.3:******For BM-Case1 and BM-Case2 with a UE-side AI/ML model, when model-based LCM is applicable at the UE, further study the level of awareness/control that NW is having on model-level LCM.***   * ***Note: UE may do the performance monitoring for model-level LCM***   Mod: Not sure what’s the additional information of this proposal compared to previous agreement |
| Google | In our view, model ID based LCM is for two-side model. So we do not support the proposal.  Mod: “***is applicable***” is included in the main bullet. If AI 9.2.1 will have an agreement not to support it, then it is not applicable |
| H3C | Suggest to make this proposal after model-ID-based LCM is agreed.  Mod: Please see the reply to Google |
| LG | Like Google, we also don’t see model-ID-based LCM is applicable to BM-Case1 and BM-Case2. Suggest to deprioritize this topic.  Mod: Please see the reply to Google |
| vivo | We support this proposal. We don’t see the difference from framework perspective for this proposal and the last proposal. Both functionality based LCM and model ID based LCM are in the same level of discussion in 9.2.1. Then we don’t see why companies can agree on the last one but not this one. Both just elaborate more details for functionality based LCM or model ID based LCM, while whether either one or both of them are supported will discussed in 9.2.1, which is orthogonal with this proposal. |
| HW/HiSi | We think there is no need to tangle with the discussion in 9.2.1, but isn’t this already covered by existing agreements?  Mod: Compared to the agreements of RAN1#110bis-e (Alt.1, Alt.2, Alt.3), the new agreements of the last meeting (NW side performance monitoring, UE side performance monitoring) provide some new details |
| CAICT | It is better to wait more conclusions from 9.2.1.  Mod: “***is applicable***” is included in the main bullet. If AI 9.2.1 will have an agreement not to support it, then it is not applicable. If we postpone the discussion to wait for progress of 9.2.1, I am afraid that there will no time left for BM-specific discussion for this issue. |
| CATT | For UE-side performance monitoring, the acknowledgement mechanism from the network is necessary.  Mod: it was discussed in GTW session. However, it is not included in the final agreement. |
| CMCC | To some extent, this proposal overlaps with existing agreement. All three alternatives can be considered for model-ID-based LCM of a UE-side AI/ML model.  ***Agreement***  ***For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the following alternatives for model monitoring with potential down-selection:***   * ***Atl1. UE-side Model monitoring***   + ***UE monitors the performance metric(s)***   + ***UE makes decision(s) of model selection/activation/ deactivation/switching/fallback operation*** * ***Atl2. NW-side Model monitoring***   + ***NW monitors the performance metric(s)***   + ***NW makes decision(s) of model selection/activation/ deactivation/switching/ fallback operation*** * ***Alt3. Hybrid model monitoring***   + ***UE monitors the performance metric(s)***   + ***NW makes decision(s) of model selection/activation/ deactivation/switching/ fallback operation***   Mod: Agree. Compared to the agreements of RAN1#110bis-e (Alt.1, Alt.2, Alt.3), the new agreements of the last meeting (NW side performance monitoring, UE side performance monitoring) provide some new details |
| Ericsson | Share the view that we can postpone this discussion based on 9.2.1 progress.  Mod: Please see the reply to CIACT |
| Xiaomi | First we would like to clarify that does it mean the hybrid model monitoring was precluded by this proposal?  Mod: No. Alt.1/2/3 are still there. This proposal is for NW-side performance monitoring, UE-side performance monitoring.  Second, for model ID based LCM, we prefer UE-side model monitoring. |
| Samsung | Whether to support model-ID-based LCM, should be discussed in general framework agenda first. We suggest to postpone this proposal.  Mod: Please see the reply to CIACT |
| Sony | We support to further study the performance monitoring based on model-ID-based LCM and functionality-based LCM. |
| NVIDIA | Wait for more progress on model-ID-based LCM in the agenda 9.2.1.  Mod: Please see the reply to CIACT |
| Qualcomm | We do not see the need for proposals 4.3.2 and 4.3.3. Essentially for both functionality-based LCM and model ID-based LCM UE-side and NW-side performance monitoring may be applicable at different levels, so we are not aligned with such categorization.  Mod: As replied before, the issue regarding model monitoring/switching under a functionality is a separate discussion. There are some related discussions in AI 9.2.1 |
| Apple | We can have more discussions. |
| MediaTek | We would like to know the reason why Hybrid performance monitoring is not on the list.  Mod: We haven’t defined “Hybrid performance monitoring” so far. That is why Proposal 4.3.1 is suggested. |
| Mod | More discussion is needed |
| ZTE | We prefer to postpone the discussion about the model-ID based LCM.  Mod: Please see the reply to CIACT |
| InterDigital | In our view, it is premature to discuss this proposal. It’s better to discuss after deciding which type of LCM is supported. |
| Futurewei | We understand the timing issue the Mod is concerning. We support the proposal. |
| Qualcomm | No need for the proposal, in our view, at this stage. |
| LG | As commented earlier, model-based LCM is mainly for two-sided model in our view. No need this agreement for BM use cases. |
| Mod | Many companies comment this proposal is not needed. Thus, the discussion is closed. |
|  |  |

###### DP 4.3.1

There are some proposals to discuss how to determine the model/functionality fails or should switch/deactivate a model/functionality. It may or may not have spec impact. The impact depends on which entity (UE or NW) makes the final decision. Since there is a limited number of companies discussing this issue in tdocs, companies are invited to share views.

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Model/functionality Identification

In previous RAN1 meeting(s), the agreement(s)/conclusion(s) were made as below:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **RAN1#111**  Working Assumption   |  |  | | --- | --- | | Terminology | Description | | Model identification | A process/method of identifying an AI/ML model for the common understanding between the NW and the UE  Note: The process/method of model identification may or may not be applicable.  Note: Information regarding the AI/ML model may be shared during model identification. |  |  |  | | --- | --- | | Terminology | Description | | Functionality identification | A process/method of identifying an AI/ML functionality for the common understanding between the NW and the UE  Note: Information regarding the AI/ML functionality may be shared during functionality identification.  FFS: granularity of functionality |   Note: whether and how to indicate Functionality will be discussed separately.  Agreement  For UE-part/UE-side models, study the following mechanisms for LCM procedures:   * For functionality-based LCM procedure: indication of activation/deactivation/switching/fallback based on individual AI/ML functionality   + Note: UE may have one AI/ML model for the functionality, or UE may have multiple AI/ML models for the functionality.   + FFS: Whether or how to indicate functionality * For model-ID-based LCM procedure, indication of model selection/activation/deactivation/switching/fallback based on individual model IDs   **RAN1#112**  Agreement  For UE-side models and UE-part of two-sided models:   * For AI/ML functionality identification   + Reuse legacy 3GPP framework of Features as a starting point for discussion.   + UE indicates supported functionalities/functionality for a given sub-use-case.     - UE capability reporting is taken as starting point. * For AI/ML model identification   + Models are identified by model ID at the Network. UE indicates supported AI/ML models. * In functionality-based LCM   + Network indicates activation/deactivation/fallback/switching of AI/ML functionality via 3GPP signaling (e.g., RRC, MAC-CE, DCI).   + Models may not be identified at the Network, and UE may perform model-level LCM.     - Study whether and how much awareness/interaction NW should have about model-level LCM * In model-ID-based LCM, models are identified at the Network, and Network/UE may activate/deactivate/select/switch individual AI/ML models via model ID.   FFS: Relationship between functionality identification and model identification  FFS: Performance monitoring and RAN4 impact  FFS: detailed understanding on model  Agreement   * AI/ML-enabled Feature refers to a Feature where AI/ML may be used.   Agreement   * For functionality identification, there may be either one or more than one Functionalities defined within an AI/ML-enabled feature. |

|  |  |
| --- | --- |
| vivo[5] | *Proposal 15: Both model ID-based LCM and functionality-based LCM should be studied for beam management.* |
| Nokia[8] | *Proposal 1. For UE-sided BM-Case1, RAN1 shall define applicable conditions for functionalities to enable functionality-based LCM.*  *Proposal 2. For UE-sided BM-Case1, RAN1 to support at least the following applicable conditions for functionalities,*  *• Supported beam prediction mode (e.g., Top-1/2/4/8 DL Tx beam prediction)*  *• Set B conditions (e.g., Measured DL RS (SSB, CSI-RS), Measured DL RS set dimension (4, 8, 12, [16]), Measured DL RS set pattern)*  *• Set A conditions (e.g., Predicted DL RS set dimension (16, 32, 64))*  *• NW-sided performance monitoring conditions (e.g., support measurements of Predicted DL RS set (full Set A, partial Set A), Measurement periodicity (100 ms, 200 ms))*  *• Conditions on supporting ML functionalities (e.g., Max number of supported functionalities (1, 2, 4, 8,.), Delay on activating a functionality (2 ms, 4 ms), Generalization condition of functionalities (yes, no))*  *Proposal 3. For UE-sided BM-Case1, RAN1 to study the following additional applicable conditions for functionalities,*  *• Conditions for UE-sided performance monitoring*  *• Conditions for data collection*  *• Conditions for predicted L1-RSRP and other metrics*  *• Conditions for assistance info required at the UE*  *Proposal 4. For UE-sided BM-Case1, the UE reports applicable conditions for functionalities using UE capability reporting.*  *Proposal 5. For UE-sided BM-Case1, the gNB creates/configures one or more functionalities to the UE with each functionality referring to an RRC configuration that contains gNB-selected appliable conditions (according to the UE capability).*  *Proposal 9. For UE-sided BM-Case2, RAN1 shall define applicable conditions for functionalities to enable functionality-based LCM.*  *Proposal 10. For UE-sided BM-Case2, RAN1 to support at least the following applicable conditions for functionalities,*  *• Supported beam prediction mode (e.g., Top-1/2/4/8 DL Tx beam prediction)*  *• Set B conditions (e.g., Measured DL RS (SSB, CSI-RS), Measured DL RS set dimension (4, 8, 12, [16]), Measured DL RS set pattern)*  *• Set A conditions (e.g., Predicted DL RS set – number of future instances (40ms, 80ms))*  *• NW-sided performance monitoring conditions (e.g., support measurements of Predicted DL RS set (full Set A, partial Set A), Measurement periodicity (100 ms, 200 ms))*  *• Conditions on supporting ML functionalities (e.g., Max number of supported functionalities (1, 2, 4, 8,.), Delay on activating a functionality (2 ms, 4 ms), Generalization condition of functionalities (yes, no))*  *Proposal 11. For UE-sided BM-Case2, RAN1 to study the following additional applicable conditions for functionalities,*  *• Conditions for UE-sided performance monitoring*  *• Conditions for data collection*  *• Conditions for predicted L1-RSRP and other metrics*  *• Conditions for assistance info required at the UE*  *Proposal 12. For UE-sided BM-Case2, the UE reports applicable conditions for functionalities using UE capability reporting.*  *Proposal 13. For UE-sided BM-Case2, the gNB creates/configures one or more functionalities to the UE with each functionality referring to an RRC configuration that contains gNB-selected appliable conditions (according to the UE capability).* |
| CATT[9] | *Proposal 8: Regarding the model identification of BM-Case1 and BM-Case2, study the following aspects as a starting point for identification information which UE should provide to gNB:*   * *Information on model functionality, e.g., BM-Case1/BM-Case2 or DL beam pair/Tx beam prediction;* * *Information of model inputs/nominal inputs, e.g., the number and pattern(s) of DL Tx beams or beam pairs in Set B;* * *Information of model outputs/nominal outputs, e.g., the number of predicted beam and/or L1-RSRP;* * *Information on assistance information for inference, e.g., the relative beam information;* * *Information on model performance;* * *Information on concurrent use with other AI/ML models and/or non-AI/ML features;* * *Information on applicable conditions.*   *Proposal 9: Regarding the functionality identification of BM-Case1 and BM-Case2, study distinguishing different functionalities by large granularity characteristics, e.g. input type and/or output type.* |
| Xiaomi[16] | *Proposal 25: BM Case 1 and BM Case 2 can be considered as different feature.*  *Proposal 26: Different functionality can be defined for different relationship between set B and set A.*   * *Alt.1: Set A and Set B are different (Set B is NOT a subset of Set A)* * *Alt.2: Set B is a subset of Set A (Set A and Set B are not the same)* * *Alt.3: Set A and Set B are the same (for BM Case 2 only)*   *Proposal 27: UE need to indicate the number of supported predicted future time instance.*  *Proposal 28: Different functionality can be defined for different repeat window for BM Case 2.*  *Proposal 29: Study assistance information from gNB to UE for model switching between different models within a same functionality.* |
| CIACT[20] | *Proposal 3: For BM-Case1 and BM-Case2, full set for functionality identification could be defined for each AI/ML-enabled Feature. Flexible functionality reporting mechanism could be considered to allow partial elements reporting within the full set.* |
| Lenovo[26] | *Proposal 4: Study UE capability on AI/ML for beam management based on Model ID or functionality-based LCM.* |
|  |  |

###### Proposal 5.1 (Closed)

In Agenda item 9.2.1, there were intensive discussions on applicable condition(s) for each functionality. In this meeting, some tdocs suggest some details dedicated for AI-based beam management. Moderator’s understanding is that the BM-specific applicable conditions need to be clearly defined in some time. Considering the current progress in Agenda item 9.2.1 on the functionality-based LCM and general applicable conditions, moderator feels it is risky to rush into some agreement(s) with too more details for AI based beam management. Thus, the main intention of Proposal 5.1 is to encourage companies to study BM-specific applicable conditions and the list of detailed conditions are just for information. Some applicable conditions are not included in the list as it seems them are common for multiple use cases. Any progress in this issue for BM-Case1 and BM-Case1 should be compliant with the output of Agenda item 9.2.1.

The related proposals in tdocs are as below:

* Nokia: Proposal 1, 2, 3, 9, 10, 11
* CATT: Proposal 9
* Xiaomi: Proposal 26, 28
* CAICT: Proposal 3

***Proposal 5.1:******For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study beam-management-specific (BM-specific) applicable conditions for functionalities. The following applicable conditions mentioned in the tdocs can be considered in further study:***

* ***Supported beam prediction mode***
* ***Set A conditions, Set B conditions, conditions on the relationship of Set A and Set B***
* ***Conditions on repeat window for BM Case 2***
* ***Conditions on input/output type***
* ***Conditions on performance monitoring***
* ***Conditions on data collection***

|  |  |
| --- | --- |
| Company | Comments |
| Mod | The intention of this proposal is to encourage more inputs. How to move forward not only depends on the inputs in this table, but also depends on the progress on AI 9.2.1 |
| Nokia/NSB | We think a good discussion is needed on this aspect. Also agree with some conditions listed by the FL.  We had following details in our contribution and coping it below to help understanding,  **Applicable conditions for BM-Case1:**   * Support Top-K DL Tx beam prediction   + K = 1, 2, 4, [8]     - *This defines the support of predicting best-K NZP CSI-RS resources based on SSB and/or CSI-RS-based RSRP measurements.* * Set B conditions   + Measured DL RS (SSB, CSI-RS)     - *Defines support of using SSB and/or CSI-RS-based RSRP measurements.*   + Measured DL RS set dimension (4, 8, 12, [16])     - *Indicates the minimum number of NZP-CSI-RS resources that shall be measured and used by the UE for predicting best-K NZP CSI-RS resources*   + Measured DL RS set pattern (e.g., fixed, pre-configured list, random)     - *Indicates the limitations on Set B conditions* * Set A conditions   + Predicted DL RS (CSI-RS)     - *Defines support of predicting CSI-RS resources*   + Predicted DL RS set dimension (16, 32, 64)     - *Indicates the maximum number of NZP-CSI-RS resources that shall be configured as the prediction NZP-CSI-RS resource set* * NW-side performance monitoring conditions   + Support measurements of Predicted DL RS set (full Set A, partial Set A)     - * *Defines the support of measuring the NZP-CSI-RS resources that correspond to Set A.*   + Measurement periodicity (100 ms, 200 ms)     - * *Indicates the minimum periodicity when supporting NZP-CSI-RS resources that correspond to Set A.* * Conditions on supporting ML functionalities   + Max number of supported functionalities (1, 2, 4, 8,)     - * *Indicates the maximum number of functionalities (e.g., number of parameter combinations that enable ML-enabled feature) that can be configured toward the UE*   + Delay in activating a functionality (2 ms, 4 ms, .)     - * *Indicates the delay required when activating or switching a functionality*   + Generalization condition of functionalities (yes, no)     - * *Indicates that the UE supports any functionality configured considering the parameter combinations of 1-4 and can be used towards the UE without any validation of whether the functionality is applicable or not.*   **Applicable conditions for BM-Case2:**   * Support Top-K DL Tx beam prediction   + K = 1, 2, 4, [8]     - *This defines the support of predicting best-K NZP CSI-RS resources based on SSB and/or CSI-RS-based RSRP measurements.* * Set B conditions   + Measured DL RS (SSB, CSI-RS)     - *Defines support of using SSB and/or CSI-RS-based RSRP measurements.*   + Measured DL RS set dimension (4, 8, 12, [16])     - *Indicates the minimum number of NZP-CSI-RS resources that shall be measured and used by the UE for predicting best-K NZP CSI-RS resources*   + Measured DL RS set periodicity (40ms, 80ms)     - *Indicates the minimum time duration for measuring NZP-CSI-RS resources that shall be measured.*   + Measured DL RS set pattern (e.g., fixed, pre-configured list, random)     - *Indicates the limitations on Set B conditions* * Set A conditions   + Predicted DL RS (CSI-RS)     - *Defines support of predicting CSI-RS resources*   + Predicted DL RS set dimension (12,16, 32, 64)     - *Indicates the maximum number of NZP-CSI-RS resources that shall be configured as the prediction NZP-CSI-RS resource set*   + Predicted DL RS set – number of future instances (40ms, 80ms)     - *Indicates the maximum time duration (or the number of future instances compared to measurement periodicity) that the NZP-CSI-RS resources can be predicted based on Set B.* * NW-side performance monitoring conditions   + Support measurements of Predicted DL RS set (full Set A, partial Set A)     - *Defines the support of measuring the NZP-CSI-RS resources that correspond to Set A.*   + Measurement periodicity (100 ms, 200 ms)     - *Indicates the minimum periodicity when supporting NZP-CSI-RS resources that correspond to Set A.* * Conditions on supporting ML functionalities   + Max number of supported functionalities (1, 2, 4, 8,)     - *Indicates the maximum number of functionalities (e.g., number of parameter combinations that enable ML-enabled feature) that can be configured toward the UE*   + Delay in activating a functionality (2 ms, 4 ms, .)     - *Indicates the delay required when activating or switching a functionality*   + Generalization condition of functionalities (yes, no)     - *Indicates that the UE supports any functionality configured considering the parameter combinations of 1-4 and can be used towards the UE without any validation of whether the functionality is applicable or not.*   Mod: In fact, the current bullets are based on your detailed list. |
| Google | We think we can revisit this after we reach consensus on how the UE and NW maintain the same understanding on the beam grid for set A/B and predicted beams. |
| LG | Input/out type do not belong to functionality-based LCM and the meaning of ‘repeat window’ is unclear. Some revision suggested below:  ***Proposal 5.1:******For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study beam-management-specific (BM-specific) applicable conditions for functionalities. The following applicable conditions ~~mentioned in the tdocs~~ can be considered in further study:***   * ***Supported beam prediction mode*** * ***Conditions on Set A/Set B configuration ~~conditions, Set A conditions, Set B conditions, conditions on the relationship of Set A and Set B~~*** * ***~~Conditions on repeat window for BM Case 2~~*** * ***~~Conditions on input/output type~~*** * ***Conditions on performance monitoring*** * ***Conditions on data collection*** * ***Note: Others are not precluded*** |
| vivo | We think applicable conditions are valid and useful not just for functionality identification but also for model identification. We suggest the following change. Specifically, we think the simplest way to define the applicable condition is to link the inference occasions to the used data set for training, then UE can just apply the same condition for this training data set. But we are open to discuss other conditions.  ***For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study beam-management-specific (BM-specific) applicable conditions ~~for functionalities~~. The following applicable conditions mentioned in the tdocs can be considered in further study:***   * ***~~Supported~~ Applied beam prediction mode*** * ***Set A conditions, Set B conditions, conditions on the relationship of Set A and Set B*** * ***Conditions on repeat window for BM Case 2*** * ***Conditions on input/output type*** * ***Conditions on performance monitoring*** * ***Conditions on data collection*** * ***Applied data set*** |
| CAICT | We tend to have an outlook on the main parts of the functionality, such as model related, configuration related, scenario/applicable conditions related, etc. We may wait for more conclusions from 9.2.1 for further discussions. |
| CATT | How to distinguish functionalities are not clear yet. We prefer to defer the discussion until more progress is achieved in AI 9.2.1. |
| CMCC | The intention to consider conditions on performance monitoring and conditions on data collection is not clear. |
| Xiaomi | What does beam prediction mode mean? BM Case 1 and BM Case 2?  And the motivation of the last two sub-bullet “*Conditions on performance monitoring”* “*Conditions on data collection*” is not clear. |
| Samsung | In our view, this is a UE-capability-like discussion. It is premature to be discussed in SI phase. Suggest to propone the discussion in WI phase. |
| Qualcomm | We agree with the direction of the proposal in general, but current wording should be updated to better understand the proposal. In the current wording, the bullets are too generic and maybe a few examples would help elaborate the intention of the proposals. For instance, “beam prediction mode” is not very clear. |
| Panasonic | We agree the direction of the proposal and some conditions listed.  In our view, the conditions on the relationship of Set A and Set B may include the association/mapping of beams within Set A and beams within Set B. In order to represent such mapping relation, one approach is to use a beam grid. The grid can be imagined as a partitioning of a reference plane such as cell coverage area on the ground. By projecting a beam onto the reference plane, it can be determined whether the beam would appear on a given grid element. If yes, the beam is indicated in the corresponding position of the grid. |
| Mod | Moderator will update the proposal with more specific bullets later. However, it is likely that the group don’t want a long list with dozens of bullets. Thus, companies are encouraged to provide Top-K (K<3? ) important aspects/conditions so that we can have a focused discussion in this meeting. |
| Futurewei | We understand that we don’t have much time left for this SI phase. But this one does seem a bit too rush, as the definitions of Feature, Functionality, Applicable Conditions, as well as the relationship among them is not clear.  Propose to defer this until 9.2.1 has clear definitions on the terms and the mechanisms. |
| ZTE | We are fine to study beam-management-specific (BM-specific) applicable conditions for functionalities and/or models, but discussion about the detailed applicable conditions can be postponed until more progress is made in agenda 9.2.1. Then, we suggest the following revision to make it more generic.  ***Proposal 5.1:******For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study beam-management-specific (BM-specific) applicable conditions for functionalities and/or models. ~~The following applicable conditions mentioned in the tdocs can be considered in further study:~~***   * ***~~Supported beam prediction mode~~*** * ***~~Set A conditions, Set B conditions, conditions on the relationship of Set A and Set B~~*** * ***~~Conditions on repeat window for BM Case 2~~*** * ***~~Conditions on input/output type~~*** * ***~~Conditions on performance monitoring~~*** * ***~~Conditions on data collection~~*** |

# Assistance information

Assistance information may be used for AI model training, inference and/or monitoring. In previous RAN1 meeting(s), the related agreement(s)/conclusion(s) were made as below:

|  |
| --- |
| **RAN1#112**  Conclusion  Regarding the explicit assistance information from UE to network for NW-side AI/ML model, RAN1 has no consensus to support the following information   * UE location * UE moving direction * UE Rx beam shape/direction   Conclusion  Regarding the explicit assistance information from network to UE for UE-side AI/ML model, RAN1 has no consensus to support the following information   * NW-side beam shape information   + E.g., 3dB beamwidth, beam boresight directions, beam shape, Tx beam angle, etc. * Note: Other information (e.g., relative information) of Tx beam(s) preserving sensitive proprietary information is a separate discussion   + e.g., some information following the same principle of Rel-17 positioning agreement |

Some related proposals are collected in the following tables:

|  |  |
| --- | --- |
| FUTUREWEI[1] | *Observation 1: Assistance information may come with additional cost like signalling overhead, extra UE measurement overhead (including complexity, power consumption, etc.). There is usually a trade-off between performance gain and the associated overhead. On another aspect, some of the proposed assistance information may be proprietary so neither the NW nor the UE is willing to expose it to the other side, unless there is a substantial gain for exposing such information.*  *Proposal 2: When assistance information is used as input, study its performance gain vs. the standards impacts and overhead, as well as whether exposing such information is appropriate/agreeable.* |
| ZTE[4] | *Observation 5: Assistance information can be used either as part of AI/ML model input or for defining applicable scenarios/configurations of the AI/ML model.*  *Proposal 11: No matter assistance information is used for model input or defining applicable scenarios/configurations, its necessity and performance gains need to be fully evaluated first in agenda 9.2.3.1.*  *Proposal 12: The introduction of any assistance information needs to consider the proprietary/privacy information disclosure issues, overhead, and standardization efforts.* |
| Vivo[5] | *Proposal 4: At least support Tx/Rx beam angle/ID information as assistance information for performance improvement for both BM-Case1 and BM-Case2. Other assistance information can be FFS.*  *Proposal 5: For the determination/selection of assistance information,*  *• The performance, model generalization and potential specification impacts should be considered.*  *• Study how to protect sensitive proprietary/privacy information and disclose beam specific related assistance information.*  *Proposal 6: Support proprietary protection mechanism for proprietary/privacy information disclosing issue. Detailed proprietary protection mechanism can be FFS.*  *Proposal 7: Suggest to use proprietary processed assistance information as model input to address performance deterioration and sensitive proprietary information disclosure issues in both BM-Case1 and BM-Case2, where a same mapping function is maintained for training and inference.* |
| OPPO[6] | *Proposal 20: For the assistance information of BM-Case1 and BM-Case2, suggest to*  *• Justify the performance benefits if assistance information is used*  *• Identify whether the used assistance information would expose proprietary and/or privacy information of either NW-side or UE-side.* |
| Nokia[8] | *Proposal 26. For BM-Case1 and BM-Case2, assistance info considered at the input of the model may not be supported via the 3GPP signalling.* |
| LGE[18] | *Proposal #1: For the UE AI/ML model training and inference, assist information on relation/association between Set A beams and Set B beams should be provided to UE. To represent beams in Set A and/or Set B while preserving sensitive proprietary information, consider following exemplary methods.*   * *Set A beams are represented by LC coefficients of Set B beams* * *Tx beam directions are represented as ordered numbers on a 2D or 3D coordinate* |
| NVIDIA[24] | *Proposal 4: Comprehensive evaluation results showing convincing performance gains is needed to nail down the essential assistance information needed for the spatial-domain DL beam prediction.*  *Proposal 6: Comprehensive evaluation results showing convincing performance gains is needed to nail down the essential assistance information needed for the temporal DL beam prediction.* |
| NEC[28] | *Proposal 1: Support angle related information (e.g., beam angle information, UE direction/orientation information) and positioning related information (e.g., UE position) as assistance information.*  *Proposal 2: For avoiding the* *proprietary/privacy of the angle related information, study implicitly providing the assistance information (e.g., angle related information) from one side to the other side.* |
|  |  |

###### DP 6

Mod’s assessment: Some companies emphasize that any assistance information should be well justified before the decision to introduce it. Moreover, most of the detailed proposals are only suggested by one company. Thus, let’s wait for more inputs

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Spec impact of model/functionality selection, activation, deactivation, switching, and fallback operation

In previous RAN1 meeting(s), the following agreements were made:

|  |
| --- |
| **RAN1#110bis-e**  Agreement  For model selection, activation, deactivation, switching, and fallback at least for UE sided models and two-sided models, study the following mechanisms:   * Decision by the network   + Network-initiated   + UE-initiated, requested to the network * Decision by the UE   + Event-triggered as configured by the network, UE’s decision is reported to network   + UE-autonomous, UE’s decision is reported to the network   + UE-autonomous, UE’s decision is not reported to the network   FFS: for network sided models  FFS: other mechanisms  Agreement  Study the specification impact to support multiple AI models for the same functionality, at least including the following aspects:   * Procedure and assistance ignaling for the AI model switching and/or selection   FFS: Model selection refers to the selection of an AI/ML model among models for the same functionality. (Exact terminology to be discussed/defined)  Agreement (AI 9.2.1)  For model selection, activation, deactivation, switching, and fallback at least for UE sided models and two-sided models, study the following mechanisms:   * Decision by the network   + Network-initiated   + UE-initiated, requested to the network * Decision by the UE   + Event-triggered as configured by the network, UE’s decision is reported to network   + UE-autonomous, UE’s decision is reported to the network   + UE-autonomous, UE’s decision is not reported to the network   FFS: for network sided models  FFS: other mechanisms |

|  |  |
| --- | --- |
| Nokia[8] | *Proposal 8. For UE-sided BM-Case1, when the UE supports more than one functionality, the gNB shall be able to de-activate/switch one of the functionalities via dynamic signaling (e.g., MAC-CE).*  *Proposal 16. For UE-sided BM-Case2, when the UE supports more than one functionality, the gNB shall be able to select/activate one of the functionalities via dynamic signaling (e.g., MAC-CE).* |
| CATT[9] | *Proposal 16: Regarding the model monitoring for BM-Case1 and BM-Case2, study the specification impacts on the following aspects:*   * *Model update/switching/fallback procedures based on model monitoring results, including the signaling exchange between the gNB and UE;*   *Trigger condition for model update/switching/fallback.* |
| NVIDIA[24] | *Proposal 8: For AI/ML based beam prediction in spatial/time domain, study potential specification impact related to assistance signalling and procedure for model configuration, model activation/deactivation, model recovery/termination, and model selection.* |
| Lenovo[26] | *Proposal 12: Dynamic switching between AI/ML based beam prediction and non-AI/ML based beam report schemes as well as dynamic switching between different AI/ML models should be supported.* |
|  |  |
|  |  |

###### DP 7

**Mod’s assessment**: The above proposals seem quite general and applicable to all sub use cases (e.g., CSI compression, Beam prediction, positioning accuracy enhancement). Not sure what spec impact is specific to BM use cases based on the above proposals.

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |

# UE/NW Capability

The related proposals/ observations are copied as below:

|  |  |
| --- | --- |
| FUTUREWEI[1] | *Proposal 4: Regarding AI/ML-based beam management, study the standards impact, including AI/ML related UE configuration/capability reporting, which may be related to AI/ML model selection/configuration (like activation/deactivation) in case multiple trained AI/ML models are deployed, or other LCM procedures.* |
| Huawei[2] | *Proposal 16: For the data collection for model training, study how to enable the UE to measure the Set A with large number of Tx beams which may be restricted by the legacy UE capability on the maximum number of configurable RS resources.*  *Proposal 36: Study the potential specification impact for UE capability, including the following aspects as a starting point:*   * *Data collection, model training, inference latency, monitoring, models switching, model updating.* * *Details can be discussed until further progress has been made for schemes themselves and their related spec impact.*   *Proposal 37: For UE capability report of the UE-side model, study the UE report of supported configurations, including at least*   * *the number of the needed data samples for training/monitoring,* * *the supported configurations of Set A and/or Set B for model training/monitoring/inference,* * *the supported values of Top-K for inference.* |
| OPPO[6] | *Proposal 16: For BM-Case1 and BM-Case2, consider the UE capability on AI/ML beam prediction at later stage.* |
| Nokia[8] | *Proposal 4. For UE-sided BM-Case1, the UE reports applicable conditions for functionalities using UE capability reporting.*  *Proposal 12. For UE-sided BM-Case2, the UE reports applicable conditions for functionalities using UE capability reporting.* |
| Lenovo[26] | *Proposal 4: Study UE capability on AI/ML for beam management based on Model ID or functionality-based LCM.*  *Proposal 5: Introduce AI/ML processing units concept for high efficiency AI/ML resource management.* |
|  |  |
|  |  |
|  |  |

###### DP 8

**Mod’s assessment**: Detailed UE capability can be discussed later

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |
|  |  |
|  |  |

# Other aspects of LCM / use cases

In previous RAN1 meeting(s), the agreement(s)/conclusion(s) were made as below:

|  |
| --- |
| **RAN1#110**  Agreement  Study the following aspects, including the definition of components (if needed) and necessity, in Life Cycle Management   * Data collection   + Note: This also includes associated assistance information, if applicable. * Model training * [Model registration] * Model deployment   + Note: Terminology is to be defined. ~~This includes process of compiling a trained AI/ML model and packaging it into an executable format and delivering to a target device.~~ * [Model configuration] * Model inference operation * Model selection, activation, deactivation, switching, and fallback operation   + ~~Note: some of them to be refined~~ * Model monitoring * Model update   + Note: Terminology is to be defined. This includes model finetuning, retraining, and re-development via online/offline training. * Model transfer * UE capability   Note: Some aspects in the list may not have specification impact.  Note: Aspects with square brackets are tentative ~~and pending terminology definition~~.  Note: More aspects may be added as study progresses.  **RAN1#110bis-e**  Agreement  Study LCM procedure on the basis that an AI/ML model has a model ID with associated information and/or model functionality at least for some AI/ML operations ~~when network needs to be aware of UE AI/ML models~~   * FFS: Detailed discussion of model ID with associated information and/or model functionality. * FFS: usage of model ID with associated information and/or model functionality based LCM procedure * FFS: whether support of model ID * FFS: the detailed applicable AI/ML operations   Agreement  Study various approaches for achieving good performance across different scenarios/configurations/sites, including   * Model generalization, i.e., using one model that is generalizable to different scenarios/configurations/sites * Model switching, i.e., switching among a group of models where each model is for a particular scenario/configuration/site   + [Models in a group of models may have varying model structures, share a common model structure, or partially share a common sub-structure. Models in a group of models may have different input/output format and/or different pre-/post-processing.] * Model update, i.e., using one model whose parameters are flexibly updated as the scenario/configuration/site that the device experiences changes over time. Fine-tuning is one example. |

The related proposals/ observations are copied as below:

|  |  |
| --- | --- |
|  |  |
|  |  |

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## Training and deployment of AI/ML model

### Training/inference at UE/NW side

In previous RAN1 meeting(s), the following agreements were made:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RAN1#109-e**  Agreement  For the sub use case BM-Case1, consider both Alt.1 and Alt.2 for further study:   * Alt.1: AI/ML inference at NW side * Alt.2: AI/ML inference at UE side   Agreement  For the sub use case BM-Case2, consider both Alt.1 and Alt.2 for further study:   * Alt.1: AI/ML inference at NW side * Alt.2: AI/ML inference at UE side   **RAN1#110**  Agreement  At least for the sub use case BM-Case1 and BM-Case2, support both Alt.1 and Alt.2 for the study of AI/ML model training:   * Alt.1: AI/ML model training at NW side; * Alt.2: AI/ML model training at UE side.   Note: Whether it is online or offline training is a separate discussion.  Working Assumption  Include the following into a working list of terminologies to be used for RAN1 AI/ML air interface SI discussion.   |  |  | | --- | --- | | Terminology | Description | | AI/ML model delivery | A generic term referring to delivery of an AI/ML model from one entity to another entity in any manner.  Note: An entity could mean a network node/function (e.g., gNB, LMF, etc.), UE, proprietary server, etc. |   **RAN1#111**  Agreement  For the sub use case BM-Case1 and BM-Case2, at least support Alt.1 and Alt.2 for AI/ML model training and inference for further study:   * Alt.1. AI/ML model training and inference at NW side * Alt.2. AI/ML model training and inference at UE side * The discussion on Alt.3 for BM-Case1 and BM-Case2 is dependent on the conclusion/agreement of Agenda item 9.2.1 of RAN1 and/or RAN2 on whether to support model transfer for UE-side AI/ML model or not   + Alt.3. AI/ML model training at NW side, AI/ML model inference at UE side |

The related proposals/observations from the contributions are copied as below:

|  |  |
| --- | --- |
| vivo[5] | *Observation 1: Report overhead may increase dramatically but with less specification impacts for Alt. 1 with enhanced beam pair prediction solution and DL Tx beam prediction solution.*  *Observation 2: Report overhead can be reduced to top-k L1-RSRP and its related Rx beam information, but assistance information including NW-side information, such as antenna configuration, Tx beam angle, etc., should be signaled to UE for Alt.2.*  *Observation 3: Due to UE side model training, if mismatch NW-side beam information is signaled to UE, significant performance deterioration can be observed for AI based beam prediction scheme in Alt.2.*  *Observation 4: Report overhead and UE energy/complexity is limited for Alt.3, but model transfer is needed.*  *Observation 5: For Alt.3, a cell specific AI solution can be achieved with generalization consideration and infra vendor may not need to disclose NW-side information such as antenna configuration, Tx beam angle, etc.*  *Proposal 2: For the sub use case BM-Case1 and BM-Case2, support to study Alt.3 for AI/ML model training and inference:*  * Alt.3. AI/ML model training at NW side, AI/ML model inference at UE side*  *Proposal 24: For Alt.3. which is AI/ML model training at NW side and inference at UE side, it has similar data collection procedure and potential specification impacts as Alt.1, i.e. both model training and model inference at NW side, for enhanced beam pair prediction and DL Tx beam prediction scheme.*  *Proposal 36: In model inference procedure, Alt.3, i.e. model training at NW side and model inference at UE side, with enhanced beam pair prediction and DL Tx beam prediction scheme has similar specification impacts as an AI model trained and inferenced at UE side.*  *Proposal 37: Study signaling aspects enhancement related to the procedure of model transfer, model registration and model activation, for the case with AI/ML model training at NW side and AI/ML model inference at UE side.* |
| OPPO[6] | *Observation 1: For BM-Case1, deploying AI/ML inference at UE side can avoid beam reporting on Set B, therefore resulting in minimum standard impact.*  *Observation 2: For BM-Case2, deploying AI/ML inference at UE side seems more reasonable, otherwise (inference at NW side) there could be overwhelming beam reporting on Set B when Set B is the same as Set A.*  *Observation 3: Whether to study the case that AI/ML model for BM-Case1 and BM-Case2 trained at NW side and delivered to UE side, highly depends on the outcome of model transfer issue discussed in agenda item 9.2.1.* |
| CATT[9] | *Proposal 3: For DL beam pair prediction with model training at NW side and inference at UE side, study how to align the mapping rule for Tx beam ID and the mapping rule for Rx beam ID between the NW and UE.* |
| IDC[11] | *Observation 3: AI/ML inference/training at NW side (Alt.1) could be a good implementation option as UE implementation is generally limited due to computational power and battery consumption than gNB implementation. However, AI/ML inference/training generally requires more detailed explicit information which leads significant reporting overhead.*  *Observation 4: AI/ML inference/training at UE side (Alt.2) can be limited due to limited computational power and battery consumption at UE implementation, however, UE can easily utilize more information that the UE acquired by measuring SSB/CSI-RS without consuming any reporting overhead.*  *Proposal 2: Support both AI/ML inference/training at NW side (Alt.1) and UE side (Alt.2) for both BM-Case1 and BM-Case2.* |
| Apple[25] | *Proposal 2:*   * *For Model training at the NW side & inference at the NW side (Alt. 1), study efficient signalling of set B selection or beam selection and RSRP representation.* * *For Model training at the NW side & inference at the UE side (Alt. 3), study model generalization performance, study model transfer/model delivery for cell-specific AI models and non cell-specific AI models.*   *Observation 2:*   * *Alt. 1 (NW side training/NW side inference) does not require disclosure of network implementation information.* * *Alt. 3 (NW side training/UE side inference) can ensure AI/ML performance when assistance information from network is not available at the UE/UE server.* |
|  |  |
|  |  |

**Mod’s assessment:** Alt.3 is depending on the discussion of model transfers in AI 9.2.1 and other WG(s) (e.g., RAN2). Let’s wait for more progress before we reopen the discussion.

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

### Online/offline training

In previous RAN1 meeting(s), the agreement(s)/conclusion(s) were made as below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **RAN1#110**  Working Assumption   |  |  | | --- | --- | | Terminology | Description | | Online training | An AI/ML training process where the model being used for inference) is (typically continuously) trained in (near) real-time with the arrival of new training samples.  Note: the notion of (near) real-time vs. non real-time is context-dependent and is relative to the inference time-scale.  Note: This definition only serves as a guidance. There may be cases that may not exactly conform to this definition but could still be categorized as online training by commonly accepted conventions.  Note: Fine-tuning/re-training may be done via online or offline training. (This note could be removed when we define the term fine-tuning.) | | Offline training | An AI/ML training process where the model is trained based on collected dataset, and where the trained model is later used or delivered for inference.  Note: This definition only serves as a guidance. There may be cases that may not exactly conform to this definition but could still be categorized as offline training by commonly accepted conventions. |   Note: It is encouraged for the 3gpp discussion to proceed without waiting for online/offline training terminologies. |

The related proposals/observations from the contributions are copied as below:

|  |  |
| --- | --- |
| OPPO[6] | *Proposal 1: For AI/ML beam prediction, starting from offline model training at current stage.* |
| IDC[11] | *Observation 20: Investigating both offline training and online training in Rel-18 is not achievable given the limited timeline for AI/ML study.*  *Observation 21: While online training requires more complicated procedures to support training in (near) real-time, offline training requires relatively simpler procedures as offline training is done by using already collected data sets.*  *Proposal 24: Prioritize offline training for the sub use case BM-Case 1 and BM-Case 2.* |
| NVIDIA[24] | *Observation 1: Offline training may be more feasible for the near future. But in the long run, it is vital that the AI/ML models can learn continuously to adapt to varying environments, site-specific conditions, and heterogenous configurations.*  *Proposal 2: For the sub use case BM-Case1 and BM-Case2, support both Alt.1 and Alt.2 for the study of AI/ML model training:*   * *Alt.1: offline training* * *Alt.2: online training* |
| Qualcomm[27] | *Proposal 1*  *For the sub use case BM-Case1 and BM-Case2 and for UE-side AI/ML models, Agenda item 9.2.3.2 should focus on offline training scenario, in which the development and training of the AI/ML model happens offline without the need to involve 3gpp signaling* |
|  |  |

Based on previous discussions and the tdocs submitted to this meeting, all companies support offline training. The controversial point is whether to support online training or not**.** Meanwhile, some companies think whether online and offline training are up to implementation and can be transparent to the other side from the perspective of 3GPP specification. Thus, no consensus on the support of online training is achieved so far.

**Mod’s assessment:** By going through all the tdocs, moderator failed to find detailed proposal(s) for specification enhancement dedicated to online training. We don’t need to discuss whether online training is supported or not if no company proposes any specific enhancement dedicated for online training. Thus, we can wait for more inputs on the enhancement dedicated to online training before we come back to this issue.

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## Model transfer

|  |  |
| --- | --- |
| Vivo[9] | *Proposal 13: For case 1 and case 2 of beam management, both collaboration level-y, and collaboration level-z can be considered.*  *Proposal 14: Take the following supportable model update choices as one aspect for defining model update levels of beam management.*  *- Choice 0: No model update during lifecycle management*  *- Choice 1: Updating model parameters w/o model transfer*  *- Choice 2: Updating model parameters with model transfer*  *- Study the lifecycle management signaling and procedures for each of the collaboration levels and model updating choices.*  *Proposal 36: In model inference procedure, Alt.3, i.e. model training at NW side and model inference at UE side, with enhanced beam pair prediction and DL Tx beam prediction scheme has similar specification impacts as an AI model trained and inferenced at UE side.*  *Proposal 37: Study signaling aspects enhancement related to the procedure of model transfer, model registration and model activation, for the case with AI/ML model training at NW side and AI/ML model inference at UE side.* |
|  |  |

**Mod’s assessment**: The discussion on spec impacts of model transfer is deferred to wait for more progress in AI 9.2.1 and/or RAN2.

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |

## General views of sub use case

In previous meetings, the following agreements/conclusion were made as below:

|  |
| --- |
| **RAN1#109-e**  Agreement  For AI/ML-based beam management, support BM-Case1 and BM-Case2 for characterization and baseline performance evaluations   * BM-Case1: Spatial-domain DL beam prediction for Set A of beams based on measurement results of Set B of beams * BM-Case2: Temporal DL beam prediction for Set A of beams based on the historic measurement results of Set B of beams * FFS: details of BM-Case1 and BM-Case2 * FFS: other sub use cases   Note: For BM-Case1 and BM-Case2, Beams in Set A and Set B can be in the same Frequency Range |

The related proposals/observations from the contributions are copied as below:

|  |  |
| --- | --- |
| IDC[11] | *Observation 1: The agreements made in RAN1#109 and RAN1#110bis-e do not preclude the case that beams in Set A and Set B in different frequency ranges.*  *Observation 2: Supporting Set A and Set B in different frequency ranges is beneficial considering different beamwidths especially when multiple cells in different Frequency Ranges are implemented toward to an identical direction.*  *Proposal 1: BM-Case 1 and BM-Case 2 with Set A and Set B in different frequency ranges are supported as well as in a same frequency range.* |
| Sony[12] | *Proposal 2 : Support Set A and Set B in different frequency bandwidth and channel observation as the input of AI/ML model (e.g., channel matrix, CIR, etc).* |
| Google[17] | *Proposal 2: Study the input from the beam report from a group of UEs for UE-group based beam prediction, where the UEs in a group share the similar location and velocity.*  *Proposal 15: The study of AI/ML based BM should consider both FR1 and FR2.* |
| NVIDIA[24] | *Proposal 1: Beam prediction in spatial domain and beam prediction in time domain should be the focal point for studying AI/ML based algorithms for beam management.* |
|  |  |

**Mod’s assessment:** RAN plenary has confirmed the representative sub use cases. Companies are encouraged to focus on other discussion (e.g., spec impacts) rather than new cases.

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |
|  |  |

## Type of beam prediction

In previous RAN1 meeting(s), the agreement(s)/conclusion(s) were made as below:

|  |
| --- |
| **RAN1#110**  Agreement  For the sub use case BM-Case1 and BM-Case2, further study the following alternatives for the predicted beams:   * Alt.1: DL Tx beam prediction * Alt.2: DL Rx beam prediction * Alt.3: Beam pair prediction (a beam pair consists of a DL Tx beam and a corresponding DL Rx beam) * Note1: DL Rx beam prediction may or may not have spec impact   **RAN1#112**  Conclusion  For the sub use case BM-Case1 and BM-Case2, “Alt.2: DL Rx beam prediction” is deprioritized. |

The related proposals/observations from the contributions are copied as below:

|  |  |
| --- | --- |
| Huawei[2] | *Observation 5: For the beam prediction mechanisms for BM-Case1 and BM-Case2, Alt.1 (DL Tx beam prediction) is a natural replacement of the legacy P1/P2 procedure for Tx beam sweeping, and is compatible with any pattern of the Rx beams.*  *Proposal 12: The need to study spec impact for DL Tx-Rx beam pair prediction additional to DL Tx beam prediction, needs to be justified firstly in 9.2.3.1. It should be shown whether it can outperform the Tx beam prediction which also can optimize the Rx beam with non-AI/ML implementations. At least following issues should be taken into account:*   * *UE rotations and Rx beam blocking (when applicable)* * *RSRP measurement errors* * *Performance/overhead/latency* * *Complexity*   *Proposal 13: For the beam prediction mechanisms for BM-Case1 and BM-Case2 if Alt.3 (beam pair prediction) is to be further studied, it should be studied for both NW-side AI/ML model and UE-side AI/ML model symmetrically.* |
| ZTE[4] | *Observation 1: Any sampling at the Rx beam space causes severe performance loss due to the lack of spatial channel characteristics after Rx beam sampling.*  *Proposal 1: Support both DL Tx beam prediction and beam pair prediction without any further down-selection.* |
| Vivo[5] | *Proposal 1: Study the two AI-based beam prediction solutions for both BM-Case1 and BM-Case2, i.e. enhanced beam pair prediction scheme and DL Tx beam prediction scheme, and considering specification impacts with generalization aspects, such as Set B construction, supported number of Tx/Rx beams, various number of antenna configurations, etc.*  *Proposal 11: Support expected Rx beam information as the AI input as one of the solutions on NW-side beam prediction for generalization to different number of Rx beams.*  *Proposal 12: Support to further study specification impact on NW-side beam pair prediction. Consider to train sufficient number of UE locations and orientations to address the coordination system mismatch issue.* |
| OPPO[6] | *Proposal 19: For BM-Case1 and BM-Case2, support Tx beam prediction (Alt.1) and beam pair prediction (Alt.3).* |
| Spreadtrum[7] | *Proposal 4: For sub use cases BM-Case1 and BM-Case2, support Alt3 Beam pair prediction as baseline.* |
| Nokia[8] | *Proposal 20. For NW-sided BM-Case1, considering beam types of Set A/B, prioritize Alt.1: DL Tx beam prediction.*  *Proposal 21. For UE-sided BM-Case1, considering beam types of Set A/B, support Alt.1: DL Tx beam prediction and Alt.3: Beam pair prediction.*  *Proposal 24. For NW-sided BM-Case2, considering beam types of Set A/B, prioritize Alt.1: DL Tx beam prediction.*  *Proposal 25. For UE-sided BM-Case2, considering beam types of Set A/B, support Alt.1: DL Tx beam prediction and Alt.3: Beam pair prediction.* |
| Intel[10] | *Proposal 1: Beam Pair prediction (Alt-3) should be supported, at least for BM-Case 1 since it can provide large latency and measurement gains for joint P2/P3 procedure* |
| IDC[11] | *Observation 18: For Rel-15 beam management, actual mapping between DL Tx beam and UE Rx beam is totally based on UE implementation.*  *Observation 19: The implementation-based UE Rx beam selection works for Rel-15, however, UE Rx beam information is crucial to accurately predict beam qualities for AI/ML based beam prediction.*  *Proposal 23: Study benefits of specification enhancements on acquiring UE Rx beam information for DL Tx beam prediction (Alt. 1) and beam pair prediction (Alt. 3).* |
| Ericsson[14] | *Proposal 1 For DL beam pair prediction of BM-Case1 and BM-Case2 with a network-side AI/ML model, study the feasibility from the following aspects as a starting point*  *o How to generalize to different UE Rx beam shapes/directions*  *o How to generalize to different UE orientation/location*  *o Overhead in reporting TX/RX beam pairs in set B and potential assistance information*  * Note 1: Potential assistance information to achieve generalization should not disclose proprietary/privacy information.*  * Note 2: Companies are encouraged to provide evaluation results for the agenda item 9.2.3.1 addressing above aspects* |
| LGE[18] | *Proposal #13: For NW-sided model, Tx beam prediction should only be considered.* |
| Samsung[19] | *Proposal 18: For predicted beams, Alt 1 (DL Tx beam prediction) is preferred.* |
| CIACT[20] | *Proposal 1: For BM-Case1 and BM-Case2, DL Tx beam prediction at gNB side and beam pair prediction at UE side could have higher priority.* |
| Apple[25] | *Proposal 1: Deprioritize beam pair prediction in the study.* |
| Qualcomm[27] | *Proposal 3*  *For BM-Case1 and BM-Case2 prioritize the study of DL Tx beam prediction.* |
| DOCOMO[29] | *Proposal 1: Identify the practical scenario for Tx beam prediction and beam pair prediction.*  *Observation 5: In DL Tx-Rx beam pair prediction with NW side model, some mechanisms to report Rx beam ID used for beam measurements are necessary.*  *Proposal 6: If RAN1 can make the consensus that the DL Rx beam information cannot be reported to NW, DL Tx-Rx beam pair prediction with NW side model should be deprioritized due to the feasibility.* |
|  |  |

Please see the discussions in Section 3.2/3.3

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## Construction of Set A and Set B

In previous RAN1 meeting(s), the following agreements and conclusions were made:

|  |
| --- |
| **RAN1#109-e**  Conclusion  For the sub use case BM-Case1, consider the following alternatives for further study:   * Alt.1: Set B is a subset of Set A   + FFS: the number of beams in Set A and B   + FFS: how to determine Set B out of the beams in Set A (e.g., fixed pattern, random pattern, …) * Alt.2: Set A and Set B are different (e.g. Set A consists of narrow beams and Set B consists of wide beams)   + FFS: the number of beams in Set A and B   + FFS: QCL relation between beams in Set A and beams in Set B   + ~~FFS: construction of Set B (e.g., regular pre-defined codebook, codebook other than regular pre-defined one)~~ * Note1: Set A is for DL beam prediction and Set B is for DL beam measurement. * Note2: The narrow and wide beam terminology is for SI discussion only and have no specification impact * Note3: The codebook constructions of Set A and Set B can be clarified by the companies.   Conclusion  For the sub use case BM-Case2, further study the following alternatives with potential down-selection:   * Alt.1: Set A and Set B are different (e.g. Set A consists of narrow beams and Set B consists of wide beams)   + FFS: QCL relation between beams in Set A and beams in Set B * Alt.2: Set B is a subset of Set A (Set A and Set B are not the same)   + FFS: how to determine Set B out of the beams in Set A (e.g., fixed pattern, random pattern, …) * Alt.3: Set A and Set B are the same * Note1: Predicted beam(s) are selected from Set A and measured beams used as input are selected from Set B. * Note2: It is up to companies to provide other alternative(s) * Note3: The narrow and wide beam terminology is for SI discussion only and have no specification impact   **RAN1#110**  Agreement  For the sub use case BM-Case1, support the following alternatives for further study:   * Alt.1: Set A and Set B are different (Set B is NOT a subset of Set A) * Alt.2: Set B is a subset of Set A * Note1: Set A is for DL beam prediction and Set B is for DL beam measurement. * Note2: The beam patterns of Set A and Set B can be clarified by the companies.   Agreement  For the sub use case BM-Case2, further study the following alternatives:   * Alt.1: Set A and Set B are different (Set B is NOT a subset of Set A) * Alt.2: Set B is a subset of Set A (Set A and Set B are not the same) * Alt.3: Set A and Set B are the same * Note1: The beam pattern of Set A and Set B can be clarified by the companies. |

The related proposals/observations are copied as below:

|  |  |
| --- | --- |
| Huawei[2] | *Observation 4: For the alternatives of the relationship between Set A and Set B under BM-Case 2, Alt.3 (Set A and Set B are the same)*   * *Can inflict compatibility issues with non-AI/ML-based UEs* * *Results into a large beam sweeping overhead during the observation phase* * *May cause unnecessary high interference to cells from neighbor UEs.*   *Proposal 11: For the study of the alternatives of the relationship between Set A and Set B under BM-Case 2,*   * *Prioritize the study of Alt.1 (Set A and Set B are different) and Alt.2 (Set B is a subset of Set A).* * *Alt.3 (Set A and Set B are the same) can be used as a benchmark for performance comparison in evaluations.* |
| H3C[3] | *Proposal 1: For the sub use case BM-Case1, prioritize the study of Alt.2(Set B is a subset of Set A) for further study.*  *Proposal 2: For the sub use case BM-Case2, prioritize the study of Alt.2(Set B is a subset of Set A) and Alt.3(Set A and Set B are the same) for further study.* |
| ZTE[4] | *Observation 4: The number of beams for measurement (i.e., set B) and for prediction (i.e., set A) is related to the trade-off between inference performance and RS overhead for beam measurement.*  *Proposal 5: For the beam set construction of BM-Case1, the sub-sampling-based method in Alt.2 (i.e., Set B is a subset of Set A) can be a starting point for the study of spatial domain beam prediction.*  *Proposal 6: For Alt.3 (i.e., Set A and Set B are the same) in the beam set construction of BM-Case2, it is useful to configure/transmit the RS resource set (i.e., set A/B) only in the measurement window to minimize the RS overhead for beam measurement.*  *Proposal 7: For Alt.3 (i.e., Set A and Set B are the same) in the beam set construction of BM-Case2, study flexible RS resource set and report configuration within the measurement window and prediction window, regardless of NW-side model or UE-side model.* |
| OPPO[6] | *Proposal 17: For BM-Case1, Set B is a subset of Set A.*  *Proposal 18: For BM-Case2, Set B and Set A are the same.* |
| Spreadtrum[7] | *Proposal 1: For sub use cases BM-Case1, focus on Alt1, i.e., Set B is a subset of Set A.* |
| Nokia[8] | *Proposal 19. For BM-Case1, considering the construction of Set A/B, prioritize Alt.2: Set B is a subset of Set A.*  *Proposal 23. For BM-Case2, considering the construction of Set A/B, prioritized “Set B and Set A are the same”.* |
| IDC[11] | *Observation 5: As using same beamwidth for all channels and signals is a general implementation* *within a frequency range, using a subset of Set A as Set B is a reasonable option if Set A and Set B are utilized in a same frequency range.*  *Observation 6: It is difficult to use a subset of Set A considering different beamwidths for beam management between different frequency ranges.*  *Observation 7: Utilization of wide beam information from a low frequency range has great potential as a low frequency range is more reliable and utilization of wide beam requires much less time and frequency resources for beam management.*  *Proposal 3: Support ‘Set B is a subset of Set A’ when Set A and Set B are utilized in a same frequency range for both BM-Case1 and BM-Case2.*  *Proposal 4: Support ‘Set A and Set B are different’ when Set A and Set B are utilized in different frequency ranges for both BM-Case1 and BM-Case2.*  *Proposal 5: AI/ML based beam management based on association between different frequency ranges should supported for both between FR1 and FR2-1 and between FR2-1 and FR2-2.* |
| Sony[12] | *Proposal 1 : For the relationship between Set A and Set B, support both Alt.1 and Alt.2 for BM-Case1 and BM-Case2.* |
| Xiaomi[16] | *Proposal 2: For spatial domain beam prediction, consider set B is a subset of set A with high priority.*  *Proposal 10: For temporal beam prediction, consider set B is same as set A with high priority.* |
| DOCOMO[29] | *Proposal 3: Define Set C and Set D as follows to facilitate the discussion*  *・Set C is a set of beams whose measurements are obtained by UE for model inputs (Set B ⊂ Set C)*  *・Set D is a set of beams which could be potentially measured by UE (Set C ⊂ Set D)*  *Proposal 4: Study the following scenario for the beam prediction*  *・Fixed Set C. UE measure the same beam per model inference.*  *・Variable Set C. UE may measure the different beam per model inference.* |
|  |  |

**Mod’s assessment**: Potential down-selection (if any) can be discussed when we have more progress (e.g., observations of evaluation results, different spec impacts, …). Let’s focus on the study on the potential spec impact of different alternatives in other sections.

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |
|  |  |

## Set B

In previous RAN1 meeting(s), the agreement(s)/conclusion(s) were made as below:

|  |
| --- |
| **RAN1#110**  Agreement   * Study the following options on the selection of Set B of beams (pairs)   + Option 1: Set B is fixed across training and inference     - FFS on the beams of Set B   + Option 2: Set B is variable (e.g., different beams (pairs) patterns in each report/measurement during training and/or inference)     - FFS on fixed or variable number of beams (pairs)     - FFS on the details   + Other options are not precluded.   + FFS on the number of beams (pairs) in Set B   **RAN1#110bis-e**  Agreement   * Study the following options on the selection of Set B of beams (pairs)   + Option 1: Set B is fixed across training and inference   + Option 2: Set B is variable (e.g., different beams (pairs) patterns in each time instance/report/measurement during training and/or inference), FFS:     - Opt A: Set B is changed following a set of pre-configured patterns     - Opt B: Set B is randomly changed among pre-configured patterns     - Opt C: Set B is randomly changed among Set A beams (pairs)     - The number of beams(pairs) in Set B can be fixed or variable     - Note: BM-Case1 and BM-Case2 may be considered for different option.   + Other options are not precluded.   Conclusion  For the sub use case BM-Case1 and BM-Case2, Set B is a set of beams whose measurements are taken as inputs of the AI/ML model, |

The related proposals/observations are copied as below:

|  |  |
| --- | --- |
| Huawei[2] | *Proposal 5: For the study of AI/ML model input for BM-Case 1 and BM-Case 2, consider fixed beams as a starting point.* |
| Vivo[5] | *Proposal 3: Deprioritize totally random pattern in set B scheme, and support to further study specification impact on Set B with pre-configured beam patterns and Set B selected from Set C. How to select pre-configured patterns and how to configure the number of beams in Set C can be FFS.* |
| Spreadtrum[7] | *Proposal 3: For the selection of Set B of beams (pairs).*   * *If AI/ML inference is at NW side, both option1 and option2 can be considered.* * *If AI/ML inference is at UE side, beams in Set B can be determined with a fix pattern.* |
|  |  |
|  |  |

**Mod’s assessment**: Let’s focus on the study on the potential spec impact of different alternatives in other sections. Any further down-selection (if any) can be discussed when we have more progress (e.g., observations based on evaluation results, spec impacts, …)

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |

## Input of BM-Case1 and BM-Case2

In previous RAN1 meeting(s), the agreements/conclusions were made as below:

|  |
| --- |
| **RAN1#109-e**  Conclusion  Regarding the sub use case BM-Case1, further study the following alternatives for AI/ML input:   * Alt.1: Only L1-RSRP measurement based on Set B * Alt.2: L1-RSRP measurement based on Set B and assistance information   + FFS: Assistance information. The following were mentioned by companions in the discussion:  Tx and/or Rx beam shape information (e.g., Tx and/or Rx beam pattern, Tx and/or Rx beam boresight direction (azimuth and elevation), 3dB beamwidth, etc.), expected Tx and/or Rx beam for the prediction (e.g., expected Tx and/or Rx angle, Tx and/or Rx beam ID for the prediction), UE position information, UE direction information, Tx beam usage information, UE orientation information, etc.     - Note: The provision of assistance information may be infeasible due to the concern of disclosing proprietary information to the other side. * Alt.3: CIR based on Set B * Alt.4: L1-RSRP measurement based on Set B and the corresponding DL Tx and/or Rx beam ID * Note1: It is up to companies to provide other alternative(s) including the combination of some alternatives * Note2: All the inputs are “nominal” and only for discussion purpose.   Conclusion  Regarding the sub use case BM-Case2, further study the following alternatives of measurement results for AI/ML input (for each past measurement instance):   * Alt.1: Only L1-RSRP measurement based on Set B * Alt 2: L1-RSRP measurement based on Set B and assistance information   + FFS: Assistance information. The following were mentioned by companies in the discussion:, Tx and/or Rx beam angle, position information, UE direction information, positioning-related measurement (such as Multi-RTT), expected Tx and/or Rx beam/occasion for the prediction (e.g., expected Tx and/or Rx beam angle for the prediction, expected occasions of the prediction), Tx and/or Rx beam shape information (e.g., Tx and/or Rx beam pattern, Tx and/or Rx beam boresight directions (azimuth and elevation), 3dB beamwidth, etc.) , increase ratio of L1-RSRP for best N beams, UE orientation information     - Note: The provision of assistance information may be infeasible due to the concern of disclosing proprietary information to the other side. * Alt.3: L1-RSRP measurement based on Set B and the corresponding DL Tx and/or Rx beam ID * Note1: It is up to companies to provide other alternative(s) including the combination of some alternatives * Note2: All the inputs are “nominal” and only for discussion purpose. |

The related proposals/observations are copied as below:

|  |  |
| --- | --- |
| Huawei[2] | *Proposal 1: For the remainder of the study item, do not consider further the CIR based on Set B as model input.*  *Proposal 2: For BM-Case 1 and BM-Case 2, use Alt.1 (Only L1-RSRP for Set B) as a starting point for the study on AI/ML input.*  *Proposal 3: For the study of AI/ML model input for BM-Case 1 and BM-Case 2, if Alt.2 is to be studied where the assistance information is in forms of implicit data categorization ID:*   * *The necessity and performance benefits of non-proprietary/non-privacy assistance information should be identified and evaluated firstly to justify a study of their specification impact.*   + *Note: implicit assistance signaling is expected to preserve privacy/proprietary information*   *Proposal 4: For Alt.4 for the BM-Case 1 and Alt.3 for BM-Case 2 for the AI/ML model input which are identical (using L1-RSRP for Set B and DL Tx and/or Rx beam ID):*   * *These two alternatives can be studied if benefits are justified by evaluation.* |
| ZTE[4] | *Proposal 10: For BM-Case1 and BM-Case2, to reduce standardization workload and avoid privacy/proprietary disclosure issues, the AI input and output can be focused on measured RSRP and/or beam ID.* |
| Vivo[5] | *Proposal 8: Regarding to BM-Case1 and BM-Case 2, at least prioritize following AI input information for further study on specification impact:*  *• L1-RSPR measurement based on Set B*  *• Corresponding DL Tx beam pointing angle/ID*  *• Corresponding DL Rx beam pointing angle/ID*  *• Expected Tx and/or expected Rx beam angle/ID*  *• Further discuss other information, such as Tx and/or Rx beam shape information, 3dB beam-width, etc.* |
| CATT[9] | *Proposal 4: For the sub use case BM-Case1 and BM-Case2, study the following alternatives for AI/ML input:*   * *Alt.1: Only L1-RSRP measurement based on Set B;* * *Alt.2: L1-RSRP measurement based on Set B and the corresponding DL Tx and/or Rx beam ID.*   *Proposal 5: Whether it is beneficial to take the assistance information as model input can be evaluated in AI 9.2.3.1.* |
| IDC[11] | *Observation 8: ‘Only L1-RSRP measurement based on Set B’ is not clear enough as the alternative does not provide any beam related information.*   * *If ‘Only L1-RSRP measurement based on Set B’ means that L1-RSRP measurements are provided in a fixed order, in our view, the input is not ‘Only L1-RSRP measurement based on Set B’.* * *Reporting L1-RSRP measurements in a fixed order is indicating L1-RSRP measurement with implicit beam related information.*   *Proposal 6: Companies supporting L1-RSRP values without beam ID should provide more details.*  *Observation 9: ‘L1-RSRP measurement based on Set B and the corresponding DL Tx and/or Rx beam ID’ can be a baseline option as AI/ML model can predict RSRP measurements with Tx and Rx beam IDs which are not provided.*  *Proposal 7: Support ‘L1-RSRP measurement based on Set B and the corresponding DL Tx and/or Rx beam ID’ as a baseline.*  *Proposal 8: Additional information such as TRP IDs and Panels IDs should be considered.*  *Proposal 9: ‘CIR based on Set B’ can be considered as an alternative only for beam management based on FR1 information.* |
| Xiaomi[16] | *Proposal 3: Support L1-RSRP and beam (pair) ID as AI/ML model input with high priority for variable set B.* |
| Google[17] | *Proposal 1: For spatial domain beam prediction, support Alt3 (CIR based on set B).*  *Proposal 8: For time-domain beam prediction, support to add CIR measurement based on set B as one alternative.* |
| NVIDIA[24] | *Observation 2: Evaluation results show that by using L1-RSRP measurement based on Set B of beams, the AI/ML-based algorithm can achieve performance comparable to that of exhaustive beam search in Set A of beams.*  *Proposal 3: For BM-Case 1, at least support L1-RSRP measurement based on Set B of beams as AI/ML model input.*  *Observation 3: Evaluation results show that by using historical optimal index, the AI/ML-based algorithm can satisfactorily yield optimal beam index prediction for future time instances.*  *Proposal 5: For BM-Case 2 (temporal DL beam prediction), at least support using historical optimal beam index based on Set B of beams as AI/ML model input.* |
| Lenovo[26] | *Proposal 1: Consider the following AI/ML model inputs for both UE-side and NW-side AI/ML inference*  * measured L1-RSRPs corresponding to all the beams within the measurement beam set B with a specific Rx beam are taken as model input for Tx beam ID prediction*  * measured L1-RSRPs corresponding to all the beams pairs which are determined by all the beams within measurement beam set B and all the UE’s Rx beam are taken as model input at least for beam pair prediction* |
|  |  |
|  |  |

**Mod’s assessment**: Different companies have different preferences on the alternatives of AI model inputs. We can wait for more progress (e.g., spec impacts, evaluation results) and then discuss whether any down-selection is needed or not if needed. Let’s focus on the study on the potential spec impact of different alternatives in other sections.

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |

## Output of BM-Case1 and BM-Case2

In previous RAN1 meeting(s), the agreement(s)/conclusion(s) are made as below:

|  |
| --- |
| **RAN1#110**  Agreement  Regarding the sub use case BM-Case1 and BM-Case2, study the following alternatives for AI/ML output:   * Alt.1: Tx and/or Rx Beam ID(s) and/or the predicted L1-RSRP of the N predicted DL Tx and/or Rx beams   + E.g., N predicted beams can be the top-N predicted beams * Alt.2: Tx and/or Rx Beam ID(s) of the N predicted DL Tx and/or Rx beams and other information   + FFS: other information (e.g., probability for the beam to be the best beam, the associated confidence, beam application time/dwelling time, Predicted Beam failure)   + E.g., N predicted beams can be the top-N predicted beams * Alt.3: Tx and/or Rx Beam angle(s) and/or the predicted L1-RSRP of the N predicted DL Tx and/or Rx beams   + E.g., N predicted beams can be the top-N predicted beams   + FFS: details of Beam angle(s) * FFS: how to select the N DL Tx and/or Rx beams (e.g., L1-RSRP higher than a threshold, a sum probability of being the best beams higher than a threshold, RSRP corresponding to the expected Tx and/or Rx beam direction(s)) * Note1: It is up to companies to provide other alternative(s) * Note2: Beam ID is only used for discussion purpose * Note3: All the outputs are “nominal” and only for discussion purpose * Note4: Values of N is up to each company. * Note5: All of the outputs in the above alternatives may vary based on whether the AI/ML model inference is at UE side or gNB side. * Note 6: The Top-N beam IDs might have been derived via post-processing of the ML-model output |

The related proposals/observations are copied as below:

|  |  |
| --- | --- |
| FUTUREWEI[1] | *Observation 2: Model outputs are typically used internally and hence without standards impact. Therefore, unless there are standards impacts involved, exact model outputs don’t need to be explicitly specified in the standards.*  *Proposal 3: Specify exact model outputs only when standards impact is involved while companies are encouraged to share their model output details for AI/ML based beam management.* |
| Huawei[2] | *Proposal 6: For BM-Case1 and BM-Case2, consider Alt. 1 as the baseline for the assumption on the AI/ML model output with changes:*   * *Alt.1: Tx and/or Rx Beam ID(s) (including the probability for the beam to be the best beam) and/or the predicted L1-RSRP of the N predicted DL Tx and/or Rx beams*    + *E.g., N predicted beams can be the Top-N predicted beams* |
| ZTE[4] | *Proposal 10: For BM-Case1 and BM-Case2, to reduce standardization workload and avoid privacy/proprietary disclosure issues, the AI input and output can be focused on measured RSRP and/or beam ID.* |
| Vivo[5] | *Proposal 9: Support to prioritize following AI output for further study on specification impact:*  * Tx and/or Rx Beam ID(s)/angle(s) and/or the predicted L1-RSRP of the N predicted DL Tx and/or Rx beams.*  * The N predicted Tx/Rx beams can be produced according to the expected beam information input to the AI model*  * FFS: study global beam ID or local beam ID*  * FFS: study global beam information, e.g. global beam ID or beam angle, with minimum exposures of implementation details*  *Proposal 10: Suggest to deprioritize Alt.2, i.e. Tx and/or Rx Beam ID(s) of the N predicted DL Tx and/or Rx beams and other information, for further study specification impact.* |
| OPPO[6] | *Proposal 21: For the output of AI/ML model for BM-Case1 and BM-Case2, suggest to include at least*  *• Tx and/or Rx Beam ID(s)*  *• The predicted L1-RSRP of the predicted Top-K DL Tx and/or Rx beams*  *• Note: the above output should be extended for F time instances for BM-Case2* |
| Intel[10] | *Proposal 3: For BM-Case1 and 2, Alt-1 (Tx and/or Rx Beam ID(s) and/or the predicted L1-RSRP of the N predicted DL Tx and/or Rx beams) should be considered as the baseline use case, with potential specification impact on how beam IDs are mapped in the spatial domain.* |
| IDC[11] | *Proposal 10: Support ‘Tx and/or Rx Beam ID(s) and/or the predicted L1-RSRP of the N predicted DL Tx and/or Rx beams’ as a baseline.*  *Proposal 11: ‘Tx and/or Rx Beam ID(s) of the N predicted DL Tx and/or Rx beams and other information’ can be considered with LOS probability.*  *Proposal 12: Benefits from utilization of TX/Rx beam angles should be clarified.* |
| Sony[12] | *Observation 1 : The complexity of beam measurement can be reduced if a subset of candidate beams can be predicted by AI/ML model.*  *Proposal 3 : For the output of AI/ML, should clearly indicate the criterion associated with the predicted beam ID in BM-case1 and BM-case2, for example, sum probabilities of being the best beams higher than a threshold, maximum dwelling time, maximum RSRP, etc.* |
| Xiaomi[16] | *Proposal 8: Support Tx and/or Rx Beam ID(s) and/or the predicted L1-RSRP of the N predicted DL Tx and/or Rx beams as the AI/ML model output with high priority.* |
| Google[17] | *Proposal 3: For spatial domain beam prediction, support the best beam possibility for each beam in Set A as the output.*  *Proposal 4: For spatial-domain beam prediction, the output for Alt3 can be the channel eigenvector used for network beam generation.*  *Proposal 9: For time-domain beam prediction, support the best beam possibility for each beam in Set A as the output.*  *Proposal 10: When AI/ML model is implemented in the NW side, the output for the AI/ML for time domain beam prediction with spec impact should be the reference angle for DL Rx beam refinement (Alt3).*  *Proposal 11: When AI/ML model is implemented in the UE side, the output for the AI/ML model for time domain beam prediction with spec impact should be the reference angle for DL Tx beam refinement (Alt3).* |
| Samsung[19] | *Proposal 17: For AI/ML output for beam prediction, Alt 1 (e.g., Tx and/or Rx Beam ID(s)) is preferred.* |
| Lenovo[26] | *Proposal 2: Support Alt 1 and Alt 2 as the AI/ML Model output for both UE-side and NW-side inference.*  *Proposal 3: When specifying the AI/ML model output, we should consider that it may be used for model monitoring.* |
| NEC[28] | *Proposal 3: Support selecting Top-N1 DL Tx and/or Rx beams according to some pre-defined rules, e.g., a sum probability of being the best beam higher than a threshold, L1-RSRP higher than a threshold.* |
|  |  |
|  |  |

**Mod’s assessment**: We will focus on the spec impact (if any) of AI model output in other section(s).

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## Misc

|  |  |
| --- | --- |
| FUTUREWEI[1] | *Proposal 5: Study Standards impact, if any, related to supporting model generalization across scenarios and/or configurations, for example, indication of a configuration change that may require additional pre-/post-processing or applying adaptation techniques.* |
| OPPO[6] | *Proposal 22: For BM-Case1 and BM-Case2, study enhancement on generalization of AI/ML model (if necessary) under heterogeneous scenarios and different Tx and/or Rx beam configurations.* |
| IDC[11] | *Observation 16: The current NR specification does not consider association between beams with different beam widths.*  *Observation 17: Utilizing association between beams with different beam widths can provide benefits for prediction accuracy e.g., robust estimation/identification of whole spatial characteristics with wide beams and accurate beam identification with narrow beams.*  *Proposal 22: Study benefits of specification enhancements on association between beams with different beam widths.* |
|  |  |
|  |  |

|  |  |
| --- | --- |
| Company | Comments |
|  |  |
|  |  |

# Summary of Discussion

## Tuesday GTW (Week 1)

***Proposal 2.3.1:***

***Regarding the ~~training~~ data collection at UE side for UE-side AI/ML model ~~trained at UE side~~, study the potential specification impact of UE reporting to network from the following aspect***

* ***Supported/preferred configurations of DL RS transmission ~~Resources~~ ~~(e.g., Set A and/or Set B, RS resources)~~***
* ***~~The number of the needed data samples~~***
* ***Other aspect(s) is not precluded***

***Proposal 2.3.2:***

***Regarding the ~~training~~ data collection at UE side for UE-side AI/ML model ~~trained at UE side~~, study the potential RAN1 specification impact (if any) to initial/trigger data collection by considering the following options as a starting point (with potential down-selection)***

* ***Option 1: data collection initialed/triggered by configuration from NW (e.g., RS configuration, configuration ID)***
* ***Option 2: ~~the corresponding~~ RS transmission request triggered by UE, including potential operations as below***
  + ***If gNB confirms the request from UE, gNB ~~and~~ performs the corresponding RS transmission***

“RAN1” is added in the main bullet to address the concern that this issue should be discussed in RAN2.

***Proposal 3.2.2:***

***For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the following additional aspects (including the necessity) ~~to facilitate~~ for AI model inference:***

* ***Quantization of L1-RSRP measurement results for UE reporting***
* ***Beam indication of multiple future time instances in one indication for BM-Case2***
* ***Measurement reporting of multiple past time instances in one reporting instance for BM-Case2 model input***
* ***Overhead reduction for the reporting of L1-RSRP measurement results***
* ***Note: Corresponding evaluations (if any) will be discussed in Agenda item 9.2.3.1***

***Proposal 3.3.2:***

***For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact (including necessity) of AI model inference from the following additional aspects on top of previous agreements:***

* ***For BM-Case1: L1 reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) in one reporting instance***
* ***For BM-Case2: L1 Reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) for each one of N time instance(s) in one reporting instance***
  + ***FFS: values of N (e.g., fixed or variable)***
* ***FFS: How to reduce the overhead***
* ***Note: The performance gains should be justified given the increased UCI payload overhead***

## Thursday GTW (Week 1)

***Proposal 2.3.2: Regarding the data collection at UE side for UE-side AI/ML model, study the potential specification impact (if any) to initiate/trigger data collection from RAN1 point of view by considering the following options as a starting point***

* ***Option 1: data collection initiated/triggered by configuration from NW***
* ***Option 2: request from UE for data collection*** 
  + ***FFS: details, e.g., if gNB confirms ~~the request from UE~~ UE’s request for RS transmission~~,~~ ~~gNB~~ and performs the corresponding RS transmission***

***Proposal 3.3.2: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact (including necessity) of AI model inference from the following additional aspects on top of previous agreements:***

* ***For BM-Case1: L1 reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) in one reporting instance***
* ***For BM-Case2: L1 Reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) for each one of N time instance(s) in one reporting instance***
  + ***FFS: values of N (e.g., fixed or variable)***
* ***FFS: How to reduce the overhead***
* ***Note1: The performance gains should be justified given the increased UCI payload overhead***
* ***Note2: The associated L1-RSRP could be predicted or measured L1-RSRP.***

***Proposal 4.1.1: For AI/ML model monitoring for BM-Case1 and BM-Case2, at least study the following alternatives as the benchmark/reference (if applicable) for performance comparison as a starting point:***

* ***Alt.1: The best beam(s) obtained by measuring beams of a set indicated by gNB (e.g., Beams from Set A)***
* ***FFS:***
  + ***Alt.2: The best beam(s) among those used for AI/ML model inputs (e.g., Beams of Set B)***
  + ***Alt.3: The beam corresponding to some or all the indicated/activated TCI state(s)***
  + ***Alt.4: The predicted best beam(s) obtained by model output (e.g., Predicted Top-K Beams)***
  + ***Alt.5: Non-AI/ML solution, to make the decision of deactivation/fallback based on the performance comparison with the AI/ML solution being monitored.***
  + ***Alt.6: AI/ML solution subject to an inactive model, to make the decision of switching/selection based on the performance comparison with the AI/ML solution being monitored.***

***Proposal 4.3.1:******For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding ~~NW-side~~ hybrid performance monitoring, study the following aspects as a starting point including the study of necessity:***

* ***Configuration/Signaling from gNB to UE for measurement and/or reporting***
* ***UE calculates the performance metric and report it ~~reporting~~ to NW ~~(e.g., for the calculation of performance metric)~~***
* ***Indication from NW for UE to do LCM operations***
* ***Other aspect(s) is not precluded***
* ***Note1: At least the performance and reporting overhead of model monitoring mechanism should be considered***

## Monday GTW (Week 2)

***Proposal 2.2.1:***

***Regarding data collection for NW-side AI/ML model, study the following options (including the combination of options) for the contents of collected data***

* ***Opt.1: M1 L1-RSRPs (corresponding to M1 beams) optionally with the indication of beams (beam pairs) based on the measurement corresponding to a beam set (e.g., Set A, Set A+B, Set B), where M1 can be larger than 4***
  + ***FFS: the range of M1***
* ***Opt.2: M2 beam (beam pair) indices based on the measurement corresponding to a beam set (e.g., Set A), where M2 can be larger than 4***
  + ***FFS: the range of M2***
  + ***FFS: How to select the M2 beam (e.g. M2 strongest beam(s))***
* ***Note1: Data collection for model training may not have 3GPP specification impact***
* ***Note2: Overhead, UE complexity and power consumption should be considered for the above options***

***Proposal 4.3.1:***

***For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding hybrid performance monitoring, study the following aspects as a starting point including the study of necessity:***

* ***Configuration/Signaling from gNB to UE for measurement and/or reporting***
* ***UE calculates the performance metric(s) and report it to NW***
* ***Indication from NW for UE to do LCM operations***
* ***Other aspect(s) is not precluded***
* ***Note1: At least UE complexity and power consumption, performance, ~~and~~ reporting overhead, and latency of model monitoring mechanism should be considered***

***Proposal 3.3.2:***

***For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the necessity, benefit(s), and potential specification impact of AI model inference from the following additional aspects on top of previous agreements:***

* ***For BM-Case1: L1 reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) in one reporting instance***
* ***For BM-Case2: L1 Reporting of more than 4 predicted beams and the associated L1-RSRP (if applicable) for each one of N time instance(s) in one reporting instance***
  + ***FFS: values of N (e.g., fixed or variable)***
* ***FFS: How to reduce the overhead***
* ***Note1: The performance gains should be justified by considering UCI payload overhead***
* ***~~Note2: The associated L1-RSRP could be predicted or measured L1-RSRP.~~***

***Proposal 3.2.2:***

***For BM-Case1 and BM-Case2 with a network-side AI/ML model, study feasibility, necessity, benefit(s) and potential specification impact from the following additional aspects for AI model inference:***

* ***Beam indication of multiple future time instances [in one indication] for BM-Case2***
* ***Measurement reporting of multiple past time instances in one reporting instance for BM-Case2***
* ***Overhead reduction for the reporting of L1-RSRP measurement results (e.g. reporting a partial Set B, L1-RSRP quantization, compressed temporal information for BM-Case2, statistics of past measurements for BM-Case2, etc.)***
* ***Beam indication based on unmeasured/outdated source RS for BM-Case1 and BM-Case2***

***Proposal 2.2.2:***

***Regarding data collection for NW-side AI/ML model, study necessity, benefits and beam-management-specific potential specification impact from RAN1 point of view on the following additional aspects***

* ***Mechanism related to the reporting~~, e.g., RRC signaling, L1 signaling, other higher-layer mechanism~~***
* ***Additional information for content of the reporting***
  + ***FFS: Information associated with or configured for the reported data samples, e.g., timestamps, SNR, data quality, etc.***
* ***~~Signaling and/or condition(s) to trigger/stop data logging (including buffering) and/or reporting~~***
* ***Reporting overhead reduction***
* ***Note1: non-3GPP based solution is a separate issue.***
* ***Note2: The framework corresponding to higher layer(s) are up to the associated WG(s)***
* ***Note 3: Overhead, UE complexity and power consumption should be considered***

# Reference

1. R1-2302322 Discussion on other aspects of AI/ML for beam management FUTUREWEI
2. R1-2302361 Discussion on AI/ML for beam management Huawei, HiSilicon
3. R1-2302432 Discussion on other aspects of AI/ML beam management New H3C Technologies Co., Ltd.
4. R1-2302440 Discussion on other aspects for AI beam management ZTE
5. R1-2302480 Other aspects on AI/ML for beam management vivo
6. R1-2302543 Other aspects of AI/ML for beam management OPPO
7. R1-2302596 Other aspects on AI/ML for beam management Spreadtrum Communications
8. R1-2302631 Other aspects on ML for beam management Nokia, Nokia Shanghai Bell
9. R1-2302698 Discussion on AI/ML-based beam management CATT
10. R1-2302793 Other aspects on AI/ML for beam management Intel Corporation
11. R1-2302826 Discussion for other aspects on AI/ML for beam management InterDigital, Inc.
12. R1-2302843 Consideration on AI/ML for beam management Sony
13. R1-2302868 Discussion on AI/ML for beam management Panasonic
14. R1-2302883 Discussion on AI/ML for beam management Ericsson
15. R1-2302907 Discussion for specification impacts on AI/ML for beam management Fujitsu
16. R1-2302978 Potential specification impact on AI/ML for beam management xiaomi
17. R1-2303053 On Enhancement of AI/ML based Beam Management Google
18. R1-2303079 Other aspects on AI/ML for beam management LG Electronics
19. R1-2303123 Discussion on potential specification impact for beam management Samsung
20. R1-2303186 Discussions on AI-ML for Beam management CAICT
21. R1-2303196 Discussion on other aspects on AI/ML for beam management ETRI
22. R1-2303227 Discussion on other aspects on AI/ML for beam management CMCC
23. R1-2303339 Other aspects on AI/ML for beam management MediaTek Inc.
24. R1-2303438 AI and ML for beam management NVIDIA
25. R1-2303478 Discussion on other aspects of AI/ML for beam management enhancement Apple
26. R1-2303527 Further aspects of AI/ML for beam management Lenovo
27. R1-2303585 Other aspects on AI/ML for beam management Qualcomm Incorporated
28. R1-2303669 Discussion on AI/ML for beam management NEC
29. R1-2303708 Discussion on other aspects on AI/ML for beam management NTT DOCOMO, INC.

# Appendix A: Contact Information

The following information was collected in the last meeting(s). Please feel free to add/update/correct contact information if needed.

|  |  |  |
| --- | --- | --- |
| Company | Name | Email |
| Moderator | Zhihua SHI | szh@oppo.com |
| Apple | Weidong Yang | Wyang23@apple.com |
| vivo | Peng SUN | sunpeng@vivo.com |
| AT&T | Thomas Novlan | thomas\_novlan@labs.att.com |
| Futurewei | Chunhui Zhu | czhu@futurewei.com |
| Xiaomi | Mingju Li | limingju@xiaomi.com |
| Lenovo | Bingchao LIU | liubc2@lenovo.com |
| Sony | Chen SUN | Chen.sun@sony.com |
| Huawei, HiSilicon | Thorsten Schier | thorsten.schier@huawei.com |
| NEC | Zhen He | he\_zhen@nec.cn |
| LG Electronics | Jiwon Kang  Haewook Park | [jw.kang@lge.com](mailto:jw.kang@lge.com)  [haewook.park@lge.com](mailto:haewook.park@lge.com) |
| Panasonic | Quan Kuang | quan.kuang@eu.panasonic.com |
| Ericsson | Henrik Ryden | Henrik.a.ryden@ericsson.com |
| Nokia, NSB | Keeth Jayasinghe  Mihai Enescu | keeth.jayasinghe@nokia.com, mihai.enescu@nokia.com |
| CATT | Yongqiang FEI | feiyongqiang@catt.cn |
| Fujitsu | Wang Xin | wangxin@fujitsu.com |
| Samsung | Zhe Chen | tom.chenzhe@samsung.com |
| CMCC | Jiazhen Zhang | zhangjiazhen@chinamobile.com |
| NVIDIA | Xingqin Lin | xingqinl@nvidia.com |
| CAICT | Xiaofeng Liu | Liuxiaofeng1@caict.ac.cn |
| OPPO | Jianfei Cao | caojianfei@oppo.com |
| MediaTek | Gyu Bum Kyung  Yu-Jen Ku | gyubum.kyung@mediatek.com  yu-jen.ku@mediatek.com |
| Intel | Avik Sengupta | avik.sengupta@intel.com |
| NTT DOCOMO | Haruhi Echigo | haruhi.echigo.fw@nttdocomo.com |
| Beijing Jiaotong University (BJTU) | Wei Chen | weich@bjtu.edu.cn |
| ZTE, Sanechips | Wenfeng Liu | liu.wenfeng@zte.com.cn |
| InterDigital | Youngwoo Kwak | youngwoo.kwak@interdigital.com |
| Qualcomm | Hamed Pezeshki | hamedp@qti.qualcomm.com |
| Spreadtrum | Dawei Ma | dawei.ma@unisoc.com |
| Charter Communications | Dumitru M. Ionescu  Samer Henry | dumitru.ionescu@charter.com  C-Samer.Henry@charter.com |
|  |  |  |

# Appendix B: Agreements

## RAN1#112

Conclusion

For the sub use case BM-Case1 and BM-Case2, “Alt.2: DL Rx beam prediction” is deprioritized.

Agreement

For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the necessity, feasibility and the potential specification impact (if needed) of the following information reported from UE to network:

* Predicted L1-RSRP(s) corresponding to the DL Tx beam(s) or beam pair(s)
  + Whether/how to differentiate predicted L1-RSRP and measured L1-RSRP
* Confidence/probability information related to the output of AI/ML model inference (e.g., predicted beams)
  + FFS: Definition/content of confidence/probability information
* Note: At least the performance and spec impact should be considered

Agreement

For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study potential specification impact of AI model inference from the following additional aspects on top of previous agreements:

* Indication of the associated Set A from network to UE, e.g., association/mapping of beams within Set A and beams within Set B if applicable
* Beam indication from network for UE reception
* Note: The second bullet may or may not have additional specification impact (e.g., legacy mechanism may be reused).

Conclusion

Regarding the explicit assistance information from UE to network for NW-side AI/ML model, RAN1 has no consensus to support the following information

* UE location
* UE moving direction
* UE Rx beam shape/direction

Conclusion

Regarding the explicit assistance information from network to UE for UE-side AI/ML model, RAN1 has no consensus to support the following information

* NW-side beam shape information
  + E.g., 3dB beamwidth, beam boresight directions, beam shape, Tx beam angle, etc.
* Note: Other information (e.g., relative information) of Tx beam(s) preserving sensitive proprietary information is a separate discussion
  + e.g., some information following the same principle of Rel-17 positioning agreement

Agreement

Regarding the performance metric(s) of AI/ML model monitoring for BM-Case1 and BM-Case2, study the following alternatives (including feasibility/necessity) with potential down-selection:

* Alt.1: Beam prediction accuracy related KPIs, e.g., Top-K/1 beam prediction accuracy
* Alt.2: Link quality related KPIs, e.g., throughput, L1-RSRP, L1-SINR, hypothetical BLER
* Alt.3: Performance metric based on input/output data distribution of AI/ML
* Alt.4: The L1-RSRP difference evaluated by comparing measured RSRP and predicted RSRP
* Other alternatives are not precluded
* Note: At least the performance and spec impact should be considered

Agreement

For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding NW-side performance monitoring, study the following aspects as a starting point including the study of necessity:

* Configuration/Signaling from gNB to UE for measurement and/or reporting
* UE reporting to NW (e.g., for the calculation of performance metric)
* Indication from NW for UE to do LCM operations
* Other aspect(s) is not precluded
* Note1: At least the performance and reporting overhead of model monitoring mechanism should be considered

Agreement

For BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding UE-side performance monitoring, study the following aspects as a starting point including the study of necessity and feasibility:

* Indication/request/report from UE to gNB for performance monitoring
  + Note: The indication/request/report may be not needed in some case(s)
* Configuration/Signaling from gNB to UE for performance monitoring
* Other aspect(s) is not precluded

## RAN1#111

Agreement

For the sub use case BM-Case1 and BM-Case2, at least support Alt.1 and Alt.2 for AI/ML model training and inference for further study:

* Alt.1. AI/ML model training and inference at NW side
* Alt.2. AI/ML model training and inference at UE side
* The discussion on Alt.3 for BM-Case1 and BM-Case2 is dependent on the conclusion/agreement of Agenda item 9.2.1 of RAN1 and/or RAN2 on whether to support model transfer for UE-side AI/ML model or not
  + Alt.3. AI/ML model training at NW side, AI/ML model inference at UE side

Agreement

Regarding the data collection for AI/ML model training at UE side, study the potential specification impact considering the following additional aspects.

* Whether and how to initiate data collection
* Configurations, e.g., configuration related to set A and/or Set B, information on association/mapping of Set A and Set B
* Assistance information from Network to UE (If supported)
* Other aspect(s) is not precluded

Agreement

For BM-Case1 and BM-Case2 with a network-side AI/ML model, study potential specification impact on the following L1 reporting enhancement for AI/ML model inference

* UE to report the measurement results of more than 4 beams in one reporting instance
* Other L1 reporting enhancements can be considered

Agreement

Regarding NW-side model monitoring for a network-side AI/ML model of BM-Case1 and BM-Case2, study the necessity and the potential specification impacts from the following aspects:

* UE reporting of beam measurement(s) based on a set of beams indicated by gNB.
* Signaling, e.g., RRC-based, L1-based.
* Note: Performance and UE complexity, power consumption should be considered.

## RAN1#110bis-e

Conclusion

For AI/ML based beam management, RAN1 has no consensus to support on studying any other sub use case in addition to BM-Case1 and BM-Case2

Note: this conclusion is independent of the discussion on the alternatives of AI/ML model inputs for BM-Case1 and BM-Case2

Conclusion

For the sub use case BM-Case1 and BM-Case2, Set B is a set of beams whose measurements are taken as inputs of the AI/ML model

***Agreement***

***For BM-Case1 with a UE-side AI/ML model, study the potential specification impact of L1 signaling to report the following information of AI/ML model inference to NW***

* ***The beam(s) that is based on the output of AI/ML model inference***
* ***FFS: Predicted L1-RSRP corresponding to the beam(s)***
* ***FFS: other information***

***Agreement***

***For BM-Case2 with a UE-side AI/ML model, study the potential specification impact of L1 signaling to report the following information of AI/ML model inference to NW***

* ***The beam(s)*** ***of N future time instance(s) that is based on the output of AI/ML model inference***
  + ***FFS: value of N***
* ***FFS: Predicted L1-RSRP corresponding to the beam(s)***
* ***Information about the timestamp corresponding the reported beam(s)***
  + ***FFS: explicit or implicit***
* ***FFS: other information***

***Working Assumption***

***For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the following L1 beam reporting enhancement for AI/ML model inference***

* ***UE to report the measurement results of more than 4 beams in one reporting instance***
* ***Other L1 reporting enhancements can be considered***

***Agreement***

***For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the NW-side model monitoring:***

* ***NW monitors the performance metric(s) and makes decision(s) of model selection/activation/ deactivation/switching/ fallback operation***

***Agreement***

***Regarding NW-side model monitoring for a network-side AI/ML model of BM-Case1 and BM-Case2, study the potential specification impacts from the following aspects***

* ***Beam measurement and report for model monitoring***
* ***Note: This may or may not have specification impact.***

***Agreement***

***For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the following alternatives for model monitoring with potential down-selection:***

* ***Atl1. UE-side Model monitoring***
  + ***UE monitors the performance metric(s)***
  + ***UE makes decision(s) of model selection/activation/ deactivation/switching/fallback operation***
* ***Atl2. NW-side Model monitoring***
  + ***NW monitors the performance metric(s)***
  + ***NW makes decision(s) of model selection/activation/ deactivation/switching/ fallback operation***
* ***Alt3. Hybrid model monitoring***
  + ***UE monitors the performance metric(s)***
  + ***NW makes decision(s) of model selection/activation/ deactivation/switching/ fallback operation***

## RAN1#110

Agreement

For the sub use case BM-Case1, support the following alternatives for further study:

* Alt.1: Set A and Set B are different (Set B is NOT a subset of Set A)
* Alt.2: Set B is a subset of Set A
* Note1: Set A is for DL beam prediction and Set B is for DL beam measurement.
* Note2: The beam patterns of Set A and Set B can be clarified by the companies.

Agreement

For the sub use case BM-Case2, further study the following alternatives:

* Alt.1: Set A and Set B are different (Set B is NOT a subset of Set A)
* Alt.2: Set B is a subset of Set A (Set A and Set B are not the same)
* Alt.3: Set A and Set B are the same
* Note1: The beam pattern of Set A and Set B can be clarified by the companies.

Agreement

For the data collection for AI/ML model training (if supported), study the following aspects as a starting point for potential necessary specification impact:

* Signaling/configuration/measurement/report for data collection, e.g., signaling aspects related to assistance information (if supported), Reference signals
* Content/type of the collected data
* Other aspect(s) is not precluded

Agreement

At least for the sub use case BM-Case1 and BM-Case2, support both Alt.1 and Alt.2 for the study of AI/ML model training:

* Alt.1: AI/ML model training at NW side;
* Alt.2: AI/ML model training at UE side.

Note: Whether it is online or offline training is a separate discussion.

Agreement

For the sub use case BM-Case1 and BM-Case2, further study the following alternatives for the predicted beams:

* Alt.1: DL Tx beam prediction
* Alt.2: DL Rx beam prediction
* Alt.3: Beam pair prediction (a beam pair consists of a DL Tx beam and a corresponding DL Rx beam)
* Note1: DL Rx beam prediction may or may not have spec impact

Agreement

Regarding the model monitoring for BM-Case1 and BM-Case2, to investigate specification impacts from the following aspects

* Performance metric(s)
* Benchmark/reference for the performance comparison
* Signaling/configuration/measurement/report for model monitoring, e.g., signaling aspects related to assistance information (if supported), Reference signals
* Other aspect(s) is not precluded

Agreement

In order to facilitate the AI/ML model inference, study the following aspects as a starting point:

* Enhanced or new configurations/UE reporting/UE measurement, e.g., Enhanced or new beam measurement and/or beam reporting
* Enhanced or new signaling for measurement configuration/triggering
* Signaling of assistance information (if applicable)
* Other aspect(s) is not precluded

Agreement

Regarding the sub use case BM-Case1 and BM-Case2, study the following alternatives for AI/ML output:

* Alt.1: Tx and/or Rx Beam ID(s) and/or the predicted L1-RSRP of the N predicted DL Tx and/or Rx beams
  + E.g., N predicted beams can be the top-N predicted beams
* Alt.2: Tx and/or Rx Beam ID(s) of the N predicted DL Tx and/or Rx beams and other information
  + FFS: other information (e.g., probability for the beam to be the best beam, the associated confidence, beam application time/dwelling time, Predicted Beam failure)
  + E.g., N predicted beams can be the top-N predicted beams
* Alt.3: Tx and/or Rx Beam angle(s) and/or the predicted L1-RSRP of the N predicted DL Tx and/or Rx beams
  + E.g., N predicted beams can be the top-N predicted beams
  + FFS: details of Beam angle(s)
* FFS: how to select the N DL Tx and/or Rx beams (e.g., L1-RSRP higher than a threshold, a sum probability of being the best beams higher than a threshold, RSRP corresponding to the expected Tx and/or Rx beam direction(s))
* Note1: It is up to companies to provide other alternative(s)
* Note2: Beam ID is only used for discussion purpose
* Note3: All the outputs are “nominal” and only for discussion purpose
* Note4: Values of N is up to each company.
* Note5: All of the outputs in the above alternatives may vary based on whether the AI/ML model inference is at UE side or gNB side.
* Note 6: The Top-N beam IDs might have been derived via post-processing of the ML-model output

## RAN1#109-e

Agreement

For AI/ML-based beam management, support BM-Case1 and BM-Case2 for characterization and baseline performance evaluations

* BM-Case1: Spatial-domain DL beam prediction for Set A of beams based on measurement results of Set B of beams
* BM-Case2: Temporal DL beam prediction for Set A of beams based on the historic measurement results of Set B of beams
* FFS: details of BM-Case1 and BM-Case2
* FFS: other sub use cases

Note: For BM-Case1 and BM-Case2, Beams in Set A and Set B can be in the same Frequency Range

Agreement

Regarding the sub use case BM-Case2, the measurement results of K (K>=1) latest measurement instances are used for AI/ML model input:

* The value of K is up to companies

Agreement

Regarding the sub use case BM-Case2, AI/ML model output should be F predictions for F future time instances, where each prediction is for each time instance.

* At least F = 1
* The other value(s) of F is up to companies

Agreement

For the sub use case BM-Case1, consider both Alt.1 and Alt.2 for further study:

* Alt.1: AI/ML inference at NW side
* Alt.2: AI/ML inference at UE side

Agreement

For the sub use case BM-Case2, consider both Alt.1 and Alt.2 for further study:

* Alt.1: AI/ML inference at NW side
* Alt.2: AI/ML inference at UE side

Conclusion

For the sub use case BM-Case1, consider the following alternatives for further study:

* Alt.1: Set B is a subset of Set A
  + FFS: the number of beams in Set A and B
  + FFS: how to determine Set B out of the beams in Set A (e.g., fixed pattern, random pattern, …)
* Alt.2: Set A and Set B are different (e.g. Set A consists of narrow beams and Set B consists of wide beams)
  + FFS: the number of beams in Set A and B
  + FFS: QCL relation between beams in Set A and beams in Set B
  + ~~FFS: construction of Set B (e.g., regular pre-defined codebook, codebook other than regular pre-defined one)~~
* Note1: Set A is for DL beam prediction and Set B is for DL beam measurement.
* Note2: The narrow and wide beam terminology is for SI discussion only and have no specification impact
* Note3: The codebook constructions of Set A and Set B can be clarified by the companies.

Conclusion

For the sub use case BM-Case2, further study the following alternatives with potential down-selection:

* Alt.1: Set A and Set B are different (e.g. Set A consists of narrow beams and Set B consists of wide beams)
  + FFS: QCL relation between beams in Set A and beams in Set B
* Alt.2: Set B is a subset of Set A (Set A and Set B are not the same)
  + FFS: how to determine Set B out of the beams in Set A (e.g., fixed pattern, random pattern, …)
* Alt.3: Set A and Set B are the same
* Note1: Predicted beam(s) are selected from Set A and measured beams used as input are selected from Set B.
* Note2: It is up to companies to provide other alternative(s)
* Note3: The narrow and wide beam terminology is for SI discussion only and have no specification impact

Conclusion

Regarding the sub use case BM-Case1, further study the following alternatives for AI/ML input:

* Alt.1: Only L1-RSRP measurement based on Set B
* Alt.2: L1-RSRP measurement based on Set B and assistance information
  + FFS: Assistance information. The following were mentioned by companions in the discussion:  Tx and/or Rx beam shape information (e.g., Tx and/or Rx beam pattern, Tx and/or Rx beam boresight direction (azimuth and elevation), 3dB beamwidth, etc.), expected Tx and/or Rx beam for the prediction (e.g., expected Tx and/or Rx angle, Tx and/or Rx beam ID for the prediction), UE position information, UE direction information, Tx beam usage information, UE orientation information, etc.
    - Note: The provision of assistance information may be infeasible due to the concern of disclosing proprietary information to the other side.
* Alt.3: CIR based on Set B
* Alt.4: L1-RSRP measurement based on Set B and the corresponding DL Tx and/or Rx beam ID
* Note1: It is up to companies to provide other alternative(s) including the combination of some alternatives
* Note2: All the inputs are “nominal” and only for discussion purpose.

Conclusion

Regarding the sub use case BM-Case2, further study the following alternatives of measurement results for AI/ML input (for each past measurement instance):

* Alt.1: Only L1-RSRP measurement based on Set B
* Alt 2: L1-RSRP measurement based on Set B and assistance information
  + FFS: Assistance information. The following were mentioned by companies in the discussion:, Tx and/or Rx beam angle, position information, UE direction information, positioning-related measurement (such as Multi-RTT), expected Tx and/or Rx beam/occasion for the prediction (e.g., expected Tx and/or Rx beam angle for the prediction, expected occasions of the prediction), Tx and/or Rx beam shape information (e.g., Tx and/or Rx beam pattern, Tx and/or Rx beam boresight directions (azimuth and elevation), 3dB beamwidth, etc.) , increase ratio of L1-RSRP for best N beams, UE orientation information
    - Note: The provision of assistance information may be infeasible due to the concern of disclosing proprietary information to the other side.
* Alt.3: L1-RSRP measurement based on Set B and the corresponding DL Tx and/or Rx beam ID
* Note1: It is up to companies to provide other alternative(s) including the combination of some alternatives
* Note2: All the inputs are “nominal” and only for discussion purpose.