3GPP TSG RAN WG1 #118 R1-2405963

Maastricht, NL, 19 Aug 2024 - 23 Aug 2024

**Agenda item:**  **9.1.1**

**Source: Tejas Networks**

**Title:** **AI/ML for Beam Management**

**Document for: Discussion and Decision**

# Introduction

RAN #102 meeting approved the Rel-19 WI on AI/ML for NR Air Interface [RP-234039] and RAN #103 meeting provided the revised WID [RP-240774], based on the AI/ML techniques to NR air interface that has been studied in FS\_NR\_AIML\_Air [TR 38.843]. In this contribution, we discuss the enhancements related to AI/ML for beam management and the work item objectives related to the beam management use case are as follows,

|  |
| --- |
| **Objectives in RP-234039**  Provide specification support for the following aspects:   * AI/ML general framework for one-sided AI/ML models within the realm of what has been studied in the FS\_NR\_AIML\_Air project [RAN2]:   + Signalling and protocol aspects of Life Cycle Management (LCM) enabling functionality and model (if justified) selection, activation, deactivation, switching, fallback     - Identification related signalling is part of the above objective   + Necessary signalling/mechanism(s) for LCM to facilitate model training, inference, performance monitoring, data collection (except for the purpose of CN/OAM/OTT collection of UE-sided model training data) for both UE-sided and NW-sided models   + Signalling mechanism of applicable functionalities/models * Beam management - DL Tx beam prediction for both UE-sided model and NW-sided model, encompassing [RAN1/RAN2]:   + Spatial-domain DL Tx beam prediction for Set A of beams based on measurement results of Set B of beams (“BM-Case1”)   + Temporal DL Tx beam prediction for Set A of beams based on the historic measurement results of Set B of beams (“BM-Case2”)   + Specify necessary signalling/mechanism(s) to facilitate LCM operations specific to the Beam Management use cases, if any   + Enabling method(s) to ensure consistency between training and inference regarding NW-side additional conditions (if identified) for inference at UE   NOTE: Strive for common framework design to support both BM-Case1 and BM-Case2 |

# Measurement report for NW sided Model

In NW sided model the gNB sweeps through transmit beams in Set B to gather data. The UE reports the L1-RSRP (Layer 1 Reference Signal Received Power) measurements of these beams back to the gNB. The gNB uses an AI/ML model to predict the top-K transmit beams in Set A based on the L1-RSRP measurements. The gNB sweeps through the predicted top-K transmit beams in Set A, and the UE measures these beams to identify the best one. The UE reports the ID of the best transmit beam in Set A back to the gNB. The UE then sweeps through different receive beams to find the best receive beam that pairs with the chosen transmit beam.

In previous meeting RAN#117 the following agreement happened on NW-sided model, for inference report.

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

It is agreed that L1-RSRPs and corresponding beam information of top M beam(s) will be reported to the gNB. Top M beam(s) can be considered as within X dB gap to the largest measured value of L1-RSRP where M and X are configured by gNB. For the maximum value of M, consider Table 6.3.1.1.2-6 in 38.212. The number of bits needed for CRI and SSBRI depends on the total number of reported beams.

A table with text and numbers

Description automatically generated

For example, if 128 CSI-RS resources in total are configured by NW in RRC, which means 7 bits are needed for reporting the CRI. So, for each beam, it needs 7 bits for CRI and 7 bits for absolute RSRP. Based on the limit of max payload size of UCI on PUCCH [38.313], the maximum potential number of beams reporting on UCI will be supported up to 1706/14=121. If all beams are reported to the gNB, the CRI field is not needed. In such cases, the maximum number of beams reporting on UCI will support 1706/7 = 243. If we consider the scenario where differential RSRP is reported to the gNB with all beams, then we can support a maximum up to (1706/4 = 426) 256 beams, which is evaluated in the study item.

**Proposal 1: For inference report at NW sided model at least for BM-case 1, consider L1-RSRPs and corresponding beam information of up to M beams within X dB gap to the largest measured value of L1-RSRP.**

**Proposal 2: For inference report at NW sided model at least for BM-case 1, maximum value of M and X dB gap to the largest measured value of L1-RSRP is configured by gNB.**

**Proposal 3: Consider M = 256 as the starting point which is agreed in study item for NW-sided model inference.**

**Proposal 4:** **For NW-sided model inference, for content of beam related information, consider CRI along with L1-RSRP as a starting point.**

|  |
| --- |
| Agreement  At least for NW sided model, for the quantization of a reported L1-RSRP value at least for the report in L1 signaling, support   * Support differential L1-RSRP reporting with legacy quantization step and range   + FFS: larger quantization step(s) than the already supported legacy quantization step for differential L1-RSRP and/or for absolute L1-RSRP   + FFS: Smaller range(s) for differential L1-RSRP than the already supported legacy range |

In legacy quantization of a reported L1-RSRP value, the largest L1-RSRP is defined by a 7-bit value with 1 dB step size, and the differential L1-RSRP is defined by a 4-bit value with 2 dB step size. As differential L1-RSRP-based reporting is a common L1-RSRP reporting method, reuse legacy differential L1-RSRP-based reporting for beam reports of NW-sided models. Additionally, according to the evaluation results captured in [6.3.2, TR 38.843], a larger quantization step size, i.e., 4 dB step size for reporting differential L1-RSRP, showed only less than 5% beam prediction accuracy degradation. Hence, it is feasible to introduce a larger quantization step size for differential L1-RSRP to reduce overhead of the beam report.

**Proposal 5: At least for NW-sided model, for differential L1-RSRP reporting, support for introducing a larger quantization step size for differential L1-RSRP reporting.**

# Performance monitoring

The following agreement happened in previous meeting RAN#117 for Type 1 performance monitoring.

|  |
| --- |
| Agreement  For BM-Case1 and BM-Case2 with a UE-side AI/ML model:   * Support Type 1 performance monitoring, including the following two options:   + Option 1 (NW-side performance monitoring):     - UE sends a report to NW (for the calculation of performance metric at NW)       * Measurement results from resource set for monitoring, e.g., L1-RSRP and/or RS index is supported as the content of the report       * FFS on other contents     - The report is at least configured/triggered by NW     - Note: this may or may not have additional spec impact   + Option 2 (UE-assisted performance monitoring):     - UE calculates performance metric(s)       * FFS how to report and what to report   + FFS whether to trigger the report based on event(s) for Option 1 and/or Option 2 * FFS Type 2 performance monitoring |

In UE-assisted performance monitoring, the NW can configure dedicated resources to UE for AI/ML functionality monitoring. UE calculates performance metrics such as beam prediction accuracy and RSRP differences. Calculated performance metrics can be reported to the NW. UE can report Top-K beam prediction accuracy or L1-RSRP difference to NW. In Top-K beam prediction accuracy reporting mode, UE shall use the output of the AI/ML model, i.e., predicted Top-K beam IDs, and compare it with the ground truth value of the best Top-1 beam over the Set A beams obtained from the measurements. For the L1-RSRP difference report, the UE shall use the output of the AI/ML model predicted, i.e., Top-K beam IDs, as well as the L1-RSRPs measured (ideal L1-RSRP) over the Set A beams.

**Proposal 6: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, consider Top-K beam prediction accuracy report for UE-assisted performance monitoring.**

**Proposal 7: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, consider L1-RSRP difference report for UE-assisted performance monitoring.**

The Type 2 performance monitoring can be used for model(s) monitoring within AI/ML functionality. Where UE can make decision(s) of model selection/activation/deactivation within a functionality. UE request to gNB for performance monitoring. The UE can indicate the information of the AI/ML functionality and the performance metric to gNB. Then, gNB can send the corresponding RS to UE for performance monitoring. In another way, NW can assign AI/ML functionality and performance metrics to the UE for performance monitoring.

**Proposal 8: For Type 2 performance monitoring for UE-sided model, the UE request for performance monitoring by indicating the AI/ML functionality and the performance metric.**

**Proposal 9: For Type 2 performance monitoring NW can assign AI/ML functionality and performance metric to the UE**

# Conclusion

In this contribution, we discuss details of ML for beam management use case and have the following proposals and observations,

**Proposal 1: For inference report at NW sided model at least for BM-case 1, consider L1-RSRPs and corresponding beam information of up to M beams within X dB gap to the largest measured value of L1-RSRP.**

**Proposal 2: For inference report at NW sided model at least for BM-case 1, maximum value of M and X dB gap to the largest measured value of L1-RSRP is configured by gNB.**

**Proposal 3: Consider M = 256 as starting point which is agreed in study item for NW-sided model inference.**

**Proposal 4: For NW-sided model inference, for content of beam related information, consider CRI along with L1RSRP as a starting point.**

**Proposal 5: At least for NW-sided model, for differential L1-RSRP reporting, support for introducing a larger quantization step size for differential L1-RSRP reporting.**

**Proposal 6: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, consider Top-K beam prediction accuracy report for UE-assisted performance monitoring.**

**Proposal 7: For BM-Case1 and BM-Case2 with a UE-side AI/ML model, consider L1-RSRP difference report for UE-assisted performance monitoring.**

**Proposal 8: For Type 2 performance monitoring for UE-sided model, the UE request for performance monitoring by indicating the AI/ML functionality and the performance metric.**

**Proposal 9: For Type 2 performance monitoring NW can assign AI/ML functionality and performance metric to the UE.**