3GPP TSG RAN WG 1 Meeting #114 R1- 23xxxxx

Toulouse, France, 21 - 25 August, 2023

Title: Reply LS on Data Collection Requirements and Assumptions

**Response to:** LS R2-2306906/R1-2306388 on Data Collection Requirements and Assumptions (from RAN2)

**Release:** Release 18

**Work Item:** FS\_NR\_AIML\_air

**Source:** TSG RAN WG1

**To:** TSG RAN WG2

Cc:

Contact Person:

**Name:** Taesang Yoo

**E-mail Address:** taesangy@qti.qualcomm.com

Send any reply LS to: 3GPP Liaisons Coordinator, mailto:3GPPLiaison@etsi.org

Attachments: -

# 1 Overall description

RAN1 thanks RAN2 for the LS on Data Collection Requirements and Assumptions.

Regarding Part A: RAN2 Assumptions on data collection that require RAN1 confirmation, RAN1 has already provided reply in R1-2308730.

Regarding Part B: Aspects of data collection that require RAN1 feedback/inputs, RAN2 asked

|  |
| --- |
| **Part B: Aspects of data collection that require RAN1 feedback/inputs**To facilitate the discussion on data collection in RAN2 for further progress, RAN2 would like RAN1 to provide feedback/inputs on the following essential aspects:* Data content
* Typical data size (value or value range) of the identified data content
* Reporting type (e.g., periodic, event triggered, other) of the identified data content
* Typical latency requirement (value or value range) to transfer the identified data content

RAN2 would require RAN1 feedback/inputs on the data collection requirements per LCM purpose (i.e., model training, inference and monitoring) for each (sub)use case, and the LCM sidedness should also be considered. Besides, RAN2 would also like to know to what extent the data would / should be specified (in detail). |

Please find RAN1’s reply as follows.

Agreement

For CSI compression (For reply LS)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LCM purpose** | **Data content** | **Typical data size (per data sample)** | **Typical latency requirement** | **Notes** |
| Training | Target CSI  | See Notes 1, 2 | Relaxed | This row applies to Type 1, Type 2, and the first or second stage of described procedure of Type 3 separate training. |
| CSI Feedback | See Note 3 | Relaxed | This is for dataset delivery for the second stage of described procedure of Type 3 separate training (either from Network side to UE side, or from UE side to Network side) and forward propagation information for Type 2 training.See Note 7 |
| Gradients for CSI Feeback | No agreement | Relaxed | This is for backward propagation for Type 2 trainingSee Note 7 |
| Inference | CSI Feedback | See Note 3 | Time-critical | Can use L1 report similar to legacy CSI |
| Monitoring | Reconstructed CSI from NW to UESee Note 6 | No agreement; [expected to be similar to target CSI for monitoring] | Near-real-time | This is called “UE-sided monitoring” in RAN1. |
| Calculated performance metricsSee Note 6 | See Note 4 | Near-real-time | This is called “UE-sided monitoring” in RAN1. |
| Target CSISee Note 6 |  See Notes 1, 2 | Near-real-time | This is called “NW-sided monitoring” in RAN1. |

Note 1: Target CSI may be precoding matrix or channel matrix. RAN1’s reply for data size is based on precoding Matrix which has been more widely evaluated than channel matrix.

Note 2: Data size for target CSI depends on the format. There is no agreement on the format or necessary precision of the target CSI. Some examples based on companies’ evaluations are: eType-II format (up to ~1000 bits), eType-II-like format (~ a few 1000 bits), and float32 format (up to ~ 150K bits). The data size may also vary depending on the configuration, and the captured value indicates the order of magnitude of the typical data size per sample as a guideline.

Note 3: There is no agreement on the CSI feedback size. Values in the order of eType II payload size may be assumed (up to ~ 1000 bits) for RAN2 discussion.

Note 4: There is no agreement on the exact metric or reporting format. An example based on companies’ evaluations is: SGCS (10s of bits)

Note 5: There are no agreements on the reporting type.

Note 6: Feasibility and necessity of the monitoring schemes listed in the table are under discussion

Note 7: RAN1 has agreed to deprioritize Type 2 training over the air interface.

Note(serve as trace in session notes)

Data size for target CSI depends on the format and configuration, for examples,

* In eType-II PC 8 format, the payload size (PMI part) for rank 1, 13 subbands, 32 ports is around 300 bits.
* In eType-II PC 8 format, the payload size (PMI part) for rank 2, 19 subbands, 32 ports is around 800 bits.
* In floating point format (32 bits per sample), the precoding matrix for 1 layer, 13 subbands, 32 ports needs around 50 kilobits. This number doesn’t account for any potential compression techniques.
* In floating point format (8 bits per sample), the precoding matrix for 4 layers, 19 subbands, 32 ports needs around 40 kilobits. This number doesn’t account for any potential compression techniques.
* In floating point format (32 bits per sample), the precoding matrix for 4 layers, 19 subbands, 32 ports needs around 150 kilobits. This number doesn’t account for any potential compression techniques.

Agreement

For CSI prediction at UE side (For reply LS)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LCM purpose** | **Data content** | **Typical data size (per data sample)** | **Typical latency requirement** | **Notes** |
| Training | Target CSI in observation and prediction window | See Notes 1, 2 | Relaxed |  |
| Inference | Predicted CSI feedback (AI/ML output) | See Note 3 | Time-critical | Can use L1 report similar to legacy CSI |
| Monitoring |  ground truth (i.e., target CSI) corresponding to predicted CSI See Note 6 | See Notes 1, 2 | Near-real-time |  |
| Calculated performance metrics / Performance monitoring outputSee Note 6 | See Note 5 | Near-real-time |  |

Note 1: Target CSI may be precoding matrix or channel matrix. RAN1’s reply for data size is based on channel matrix which has been more widely evaluated than precoding Matrix.

Note 2: Data size for target CSI depends on the format. There is no agreement on the format or precision of the target CSI. The data size may also vary depending on the configuration, and the captured value indicates the order of magnitude of the typical data size per sample as a guideline. One example based on companies’ evaluations is up to around 1.5Mbits, assuming float 32 and 10 CSI-RS observation instances as input to predict one future CSI instance.

Note 3: There is no agreement on the predicted CSI feedback size. Values in the order of eType II payload size may be assumed (up to ~ 1000 bits) for RAN2 discussion.

Note 4: There are no agreements on the reporting type.

Note 5: There is no agreement on the performance metric or monitoring output details.

Note 6: Feasibility and necessity of the monitoring schemes listed in the table are under discussion.

Note (serve as trace in session notes)

Data size for target CSI depends on the format and configuration, for examples,

* In floating point format (32 bits per sample), the channel matrix for 4 layers, 19 subbands (one matrix per subband), 32 ports needs around 150 kilobits per CSI-RS instance. Assuming 10 CSI-RS observation instances as input to predict one future CSI instance, the total is around 1.5M bits. This number doesn’t account for any potential compression techniques.

## For beam management

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LCM purpose** | **UE-side/NW-side models** | **Data content** | **Typical data size (per data sample)** | **Typical latency requirement** | **Notes** |
| Training | UE-side, NW-side | L1-RSRPs [and beam-IDs] for Set BL1-RSRPs and/or beam-IDs of a subset (including one beam) or the full set of Set A | See Note 1 for L1-RSRPs | Relaxed |  |
| Inference | UE-side | Beam prediction results | Small (10s of bits) | Time-critical | RAN1 has agreed to consider L1 signalling for this reporting |
| NW-side | L1-RSRP, [Beam-ID] for Set B | See Note 1 for L1-RSRPs | Time-critical |
| Monitoring | UE-side | Event occurrence and/or calculated performance metrics See Note 4 | Small (10s of bits) | Near-real-time |  |
| UE-side | L1-RSRPs and/or beam-IDs of a subset of Set A (including one beam) See Note 4 | See Note 1 for L1-RSRPs | Near-real-time |  |
| NW-side  | L1-RSRPs and/or beam-IDs of a subset of Set A (including one beam)See Note 4 | See Note 1 for L1-RSRPs | Near-real-time |  |

Note 1: There is no agreement on the data size of L1-RSRPs for Set A or Set B, but the following typical data size is provided as guidance for RAN2 discussion. Based on existing L1-RSRP reporting methodology, i.e., 7 bits for the strongest beam and 4 bits for the remaining beams, for Set B = 16 as an example, the typical data size would be 67 (hence up to ~100 bits), and for Set A = 128 as an example, the typical data size would be 515 (hence up to ~500 bits) if all beams in Set A were to be collected. For BM Case 2, the data size L1-RSRPs for Set A and Set B represents the data size per predicted future time instance and per history measurement time instance, respectively.

Note 2: There are no agreements on the reporting type.

Note 4: Feasibility and necessity of the monitoring schemes listed in the table are under discussion.

Note 5: For BM Case 2, the typical value of the number of history measurement time instance used in evaluations is [2~8] and typical value of the number of predicted future time instance is [1~4].

## For positioning

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LCM purpose** | **Case** | **Data content** | **Typical data size (per data sample)** | **Typical latency requirement** | **Notes** |
| Training | All Cases | Measurements (corresponding to model input): timing, power, and/or phase infoSee Note 2 | Size depends on measurement type (timing, power, and/or phase info) and report format:100s bits to 1000s bits per PRS/SRS resourceSee Note 3 | Relaxed |  |
| Direct AI/ML positioning | Label: Location coordinates as model output | 56 to 144 bits See Note 3 | Relaxed |  |
| AI/ML assisted positioning | Label: Intermediate positioning measurement (timing info, [RSRP/RSRPP], LOS/NLOS indicator) as model outputSee Note 2 | 10s bits to 100s bits per PRS/SRS resourceSee Note 3 | Relaxed |  |
| Inference | 1 | Location coordinates as model output | 56 to 144 bitsSee Note 3 | See Note 5 |  |
| 2a, 3a | Intermediate positioning measurement (timing info, [RSRP/RSRPP], LOS/NLOS indicator) as model outputSee Note 2 | 10s bits to 100s bits per PRS/SRS resourceSee Note 3 | See Note 5 |  |
| 2b, 3b | Measurements (corresponding to model input):Timing, power, and/or phase info See Note 2 | Size depends on measurement type (timing, power, and/or phase info) and report format:100s bits to 1000s bits per PRS/SRS resourceSee Note 3 | See Note 5 |  |
| Monitoring | All Cases | RAN1 has studied initial listing of monitoring metrics  | RAN1 is still working on deciding data/metrics and their sizes. | Near-real-time | See Note 6, 7, and 8 |

Note 1: The necessity and feasibility of difference cases (Case1 to Case3b) needs further discussion/conclusion.

Note 2: For measurements as model input, no agreement on measurement types (i.e., time, power, and/or phase) in RAN1 for all cases (i.e., Case1 to Case3b). Measurement types (including their necessity) and sizes/dimension needs to be further discussed. Candidate measurement types discussed/evaluated for model input include CIR (contains timing, power and phase information), PDP (contains timing and power information), DP (contains timing information). For labels (i.e., model output) of AI/ML assisted positioning (Case2a, Case3a), RAN1 identified an initial listing of candidates that provide performance benefits (i.e., timing info, [RSRP/RSRPP], LOS/NLOS indicator).

Note 3: The measurement size of one data sample = (measurement data size of one PRS/SRS resource)\*(number of PRS/SRS resources needed for model input). The label size of one data sample = (label data size of one PRS/SRS resource)\*(number of PRS/SRS resources needed for model output). The quantization and bit representation of time, power, and phase information (including their necessity) still need to be further discussed. As a reference to existing timing and power representation, multipath measurement reporting for measurements of one PRS resource in the existing specification TS 37.355 is shown below for UE reporting to LMF. Similar measurement reporting exists in TS 38.455 for gNB reporting to LMF for one SRS resource:

nr-DL-PRS-RSRP-Result-r16 à 7 bits

nr-DL-PRS-RSRP-ResultDiff-r16 à 6 bits

nr-RSTD-r16 à 16 to 21 bits

nr-RelativeTimeDifference-r16 à 9 to 14 bits

A potential upper bound can be computed with timing info as 21 bits for first arrival and 14 bits for relative timing; power as 7 bits for first value and 6 bits for relative powers/real values. While a potential lower bound can be computed with timing info as 16 bits for first arrival and 9 bits for relative timing; power as 7 bits for first value and 6 bits for relative powers/real value info. Existing specification allows reporting of up to 64 PRS/SRS resources per frequency layer for one positioning fix. For evaluations, most companies considered up to 18 TRPs. It should be noted that AI/ML positioning is not restricted to work only with maximum of 18 TRPs.

* Example of calculation on a potential lower bound on measurement size per PRS/SRS resource:
	+ A potential lower bound on measurement size per PRS/SRS resource can be calculated as follows (assuming timing only for 9 measurements per PRS/SRS resource): 16 + 9\*8 = 88 bits. The total lower bound can be 88\*N bits, where N is number of PRS/SRS resources used as model input for obtaining a positioning fix.
* Example of calculation of a potential upper bound on measurement size per PRS/SRS resource:
	+ A potential upper bound on measurement size per PRS/SRS resource can be calculated as follows (assuming timing, power, and phase for 256 measurements per PRS/SRS resource and assuming 8 bit representation of each real number): 2\*(8\*256) = 4096 bits. The total upper bound can be 6665\*N bits, where N is number of PRS/SRS resources used as model input for obtaining a positioning fix.
* For location coordinates (corresponding to model output)
	+ The bit representation of location coordinates depends on the type of shape, resolution, and uncertainty used to indicate the location (e.g., ellipsoid point, ellipsoid point with uncertainty circle, high accuracy ellipsoid with uncertainty ellipsoid, etc.) as listed in TS 23.032. The range of bit representation for location coordinates can be 7 bytes to 18 bytes (i.e., 56 to 144 bits). The location information report in existing specifications may contain additional information besides location coordinates (e.g., velocity, location error, integrity info, etc.)
* For intermediate positioning measurement (corresponding to model output):

The quantization and bit representation of time, [RSRP/RSRPP], and LOS/NLOS information (including their necessity) as model output still need to be discussed in an appropriate working group. As a reference to existing timing representation in Rel17 [TS 37.355], an example on the label size can be of 21 bits per PRS/SRS resource while assuming model output produces one timing of 21 bits per PRS/SRS resource. The label size can be 21\*N bits, where N is number of PRS/SRS resources for which intermediate positioning measurement has been generated. If power info (7 bits per PRS/SRS resource) is included, the label size becomes 28\*N bits. Note 4: No agreement on reporting types (i.e., periodicity, event-triggered/on-demand, etc.). As a reference, the existing positioning procedures consider periodic and triggered/on-demand reporting. For periodic reporting, the reporting interval can be {1, 2, 4, 8, 10, 16, 20, 32, 64} seconds (see IE PeriodicalReporting [TS 37.355] and IE UEReportingInformation [TS 38.455]).

Note 5: There are no agreements on the reporting latency. For inference, as a reference, the existing positioning procedures consider different response timing depending on quality of service and target device type (e.g., target device supporting NB-IoT, HA GNSS, etc.). The response time is measured between receipt of location request and transmission of report. The response time can be between 1 and 128 timing units for regular target devices, where a one timing unit can be {ten-milli-seconds, seconds, ten-seconds} (see IE ResponseTime [TS 37.355]).

Note 6: RAN1 agreed on an initial listing of entities that can derive the monitoring metric for AI/ML positioning for different cases (Case1 to Case3b):

 -1: At least UE derives monitoring metric

 -2a: At least UE and LMF (based on ground truth) derive monitoring metric

 -3a: At least gNB/TRP and LMF (based on ground truth) derive monitoring metric

 -2b and 3b: At least LMF derives monitoring metric

Note 7: No agreement yet a monitoring decision entity or their mapping to other entities (e.g., entity running the inference, entity deriving the monitoring metric, etc.).

Note 8: Further details on monitoring data/metrics (including their feasibility and whether there are spec impacts) and mapping to different AI/ML positioning cases need further discussion. No agreement on the need or content of data collection for the purpose of model monitoring. This is to be discussed separately for Cases 1/2a/2b/3a/3b. These are descriptions on metrics that have been studied by RAN1:

 - Statistics of the difference between model output and provided (approximate) ground truth label,

 - Statistics of measurement and/or model input compared to the statistics associated with the training data

 - Statistics of model output compared to the statistics associated with the training data and/or its own previous inference output.

## Common Notes for all sub-use-cases:

* In answering latency requirements, RAN1 used the following descriptions:
	+ Relaxed (e.g., minutes, hours, days, or no latency requirement)
	+ Near-real-time (e.g., several tens of msecs to a few seconds)
	+ Time-critical (e.g., a few msecs)
* In the reply, RAN1 captured the typical data size per each data sample.
* Model training is assumed to be offline training.
* In RAN1’s answer, RAN1 did not list assistance information. RAN1 has informed RAN2 of related conclusions/agreements/observations regarding assistance information in the RAN1 response to part A.
* There may be other information useful for training at the UE-side, NW side, or neutral-side not included in the tables. For example, in positioning enhancement, some information has been considered as potential spec impact in agreement of RAN1-114 meeting (e.g., quality indicators, time stamps, [scenario identifier, LOS/NLOS condition, timing error]). As another example, for beam prediction, additional information for the contents of collected data was considered (e.g., timestamps, SNR, data quality, etc.).
* There may be configuration information that may be collected together with data. Such information is not included in the tables. Examples include RS configuration(s) for deriving measurements, configuration related to Set A and/or Set B, information on association/mapping of Set A and Set B.
* In this reply for Part B, the term 'NW-side monitoring' is not explicitly used since RAN1’s understanding of the term is not fully aligned with RAN2 terminology. It should also be noted that in the RAN1 response to part A, RAN1 used the term ‘NW-sided monitoring’ aligned with RAN2.
* For monitoring, RAN1 provided replies only for near-real-time monitoring. The requirements for data collection for relaxed monitoring, if necessary, can be considered to be similar to offline training requirements.
* This LS reply is meant to capture existing RAN1 agreements/conclusions/observations and discussions for the purpose of replying the RAN2 LS; The LS reply does not serve as additional agreements/conclusions/observations beyond what RAN1 has already agreed/concluded/observed.

# 2 Actions

To RAN2

**ACTION:** RAN1 respectively asks RAN2 to take the above information into account in RAN2’s discussions.

# 3 Dates of next RAN1 meetings

RAN1#115 13-17 November 2023 Chicago, US

== The following will NOT be included in the LS reply. ==

# 4 LS reply on Part A

(This section was copied from R1-2308730 for ease of reference.)

RAN1 thanks RAN2 for the LS on Data Collection Requirements and Assumptions.

Regarding **Part A: RAN2 Assumptions on data collection that require RAN1 confirmation,** please find RAN1’s reply as follows. It reflects the current status of RAN1 discussion.

Regarding Assumption 1 of Part A,

|  |
| --- |
| Assumption 1:RAN2 assumes that for the data collection in some scenarios (e.g., internal data up to implementation or the existing data are enough), possibly no RAN2 specification effort is needed in some scenarios, e.g. (not exhaustive):* For model inference of the UE-sided model, input data for model inference is available inside the UE.
* For UE-side (real-time) monitoring of the UE-sided model, performance metrics are available inside the UE. UE can independently monitor a model's performance without any data input from NW.
 |

RAN1 discussed the two bullets under Assumption 1 and made following clarification on them:

* For model inference of the UE-sided model, input data for model inference is available inside the UE.
* For UE-side ~~(real-time)~~ performance monitoring of the UE-sided model, in some cases, e.g., for CSI prediction and beam prediction, performance metrics are available inside the UE. UE can independently monitor a model's performance without any data input from NW.
	+ Note: RAN1’s understanding is that “data input” in the above does not include assistance information that a model may additionally use for performance metric calculation.

Regarding Assumption 2 of Part A,

|  |
| --- |
| Assumption 2:For the latency requirement of data collection, RAN2 assumes:* For all types of offline model training (i.e., UE- /NW-/ two-sided model training), there is no latency requirement for data collection
* For model inference, when required data comes from other entities, there is a latency requirement for data collection
* For (real-time) model monitoring, when required monitoring data (e.g., performance metric) comes from other entities, there is a latency requirement for data collection.
 |

RAN1 confirms Assumption 2 in RAN2 LS.

Regarding Assumption 3 of Part A,

|  |
| --- |
| Assumption 3:RAN2 assumes that the analysis/selection of the data collection frameworks should focus on the RRC\_CONNECTED state (for both data generation and reporting). Analysis and potential enhancement of the non-connected state can be revisited when needed. |

RAN1 confirms RAN2’s Assumption 3 for CSI compression, CSI prediction, beam prediction and Positioning use cases.

For positioning, it is noted that existing specification supports DL PRS measurement and UE positioning in both RRC\_CONNECTED and RRC\_INACTIVE state.

Regarding Assumption 4 of Part A,

|  |
| --- |
| Assumption 4:For the data generation entity and termination entity deployed at different entities, RAN2 made the following assumptions:* For CSI enhancement and beam management use cases:
* For model training, training data can be generated by UE/gNB and terminated at gNB/OAM/OTT server.
* For NW-sided model inference, input data can be generated by UE and terminated at gNB.
* For UE-side model inference, input data/assistance information can be generated by gNB and terminated at UE.
* For model monitoring at the NW side, performance metrics can be generated by UE and terminated at gNB.
* For positioning enhancement use case:
* For model training, training data can be generated by UE/gNB and terminated at LMF/OTT server.
* For NW-sided model inference, input data can be generated by UE/gNB and terminated at LMF and/or gNB.
* For UE-side model inference, input data/assistance information can be generated by LMF/gNB and terminated at the UE.
* For model monitoring at the NW side, performance metrics can be generated by UE/gNB and terminated at LMF.
 |

RAN1 discussed Assumption 4 of Part A for CSI compression, CSI prediction, Beam management and positioning use case separately and made following clarification for each use case based on the Assumption 4 of Part A of RAN2 LS.

* For CSI compression ~~enhancement~~ ~~and beam management~~ use case:
* For model training, training data can be generated by UE/gNB ~~and terminated at gNB/OAM~~/~~OTT server~~
* For ~~NW-sided model inference~~ NW-part of two-sided model inference, input data can be generated by UE and terminated at gNB.
* For UE-side model inference UE-part of two-sided model inference, input data is internally available at UEinput data/assistance information can be generated by gNB and terminated at UE.
* For ~~model~~ performance monitoring at the NW side, calculated performance metrics (if needed) or data needed for performance metric calculation (if needed) can be generated by UE and terminated at gNB.
* For CSI prediction enhancement and beam management use case:
* For model training, training data can be generated by UE~~/gNB and terminated at gNB/OAM/OTT server~~.
* For NW-sided model inference, input data can be generated by UE and terminated at gNB.
* For UE-side model inference, input data~~/assistance information~~ is internally available at UE ~~can be generated by gNB and terminated at UE~~.
* For performance~~model~~ monitoring at the NW side, calculated performance metrics (if needed) or data needed for performance metric calculation (if needed) can be generated by UE and terminated at gNB.
* For ~~CSI enhancement and~~ beam management use case:
* For model training, training data can be generated by UE/gNB ~~and terminated at gNB/OAM/OTT server~~.
* For NW-sided model inference, input data can be generated by UE and terminated at gNB.
* For UE-side model inference, input data~~/assistance information~~ is internally available at UE. ~~can be generated by gNB and terminated at UE.~~
* For performance~~model~~ monitoring at the NW side, calculated performance metrics (if needed) or data needed for performance metric calculation (if needed) can be generated by UE and terminated at gNB.
* For positioning enhancement use case:
* For model training, training data can be generated by UE/PRU/gNB/LMF ~~and terminated at LMF/OTT server~~.
* For LMF~~NW~~-sided model inference (Case 2b, Case 3b), input data can be generated by UE/gNB and terminated at LMF ~~gNB~~.
* For gNB-sided model inference (Case 3a), input data is internally available at gNB.
* For UE-side model inference (Case 1, Case 2a), input data~~/assistance information~~ is internally available at UE ~~can be generated by LMF/gNB and terminated at the UE~~.
* For ~~model~~performance monitoring at the ~~NW~~LMF side, calculated performance metrics (if needed) or data needed for performance metric calculation (if needed) can be generated by UE/gNB and terminated at LMF.
* For ~~model~~performance monitoring at the ~~NW~~gNB side, calculated performance metrics (if needed) or data needed for performance metric calculation (if needed) can be generated by at least gNB.

Note: In RAN1’s answer to Assumption 4, RAN1 did not reply on the different NW entities for training (gNB/CN/LMF/OAM) as it is out of RAN1’s expertise that RAN1 cannot confirm.

Note: For the above replies for Assumption 1~4 in Part A, RAN1’s understanding is that “input data” in the RAN2 LS does not include assistance information that a model may additionally use as model input. In RAN1’s answer, RAN1 did not reply on assistance information, and informs RAN2 of related conclusions/agreements/observations in the Appendix.

Regarding Part B: Aspects of data collection that require RAN1 feedback/inputs, RAN1 is yet to discuss the Part B and will reply later.

# 5 Discussions

Please provide comments either below or directly in Section 1 using MS-Word’s commenting tools.

|  |  |
| --- | --- |
| Company | Comments |
| Mod | Regarding comments to remove “Gradients for CSI Feeback” from CSI compression, please note that we list contents that move from one entity to another, regardless of whether they may or may not have spec impact. RAN1 is not making any judgement whether they have spec impact in RAN2 or other working groups.  |
| Panasonic | For CSI compression- In LCM purpose column, "Inference" is started from capital but "tranining" and "monitoring is not from capital. Alighment among them is necessary like "Training" and "Monitoring". - "if needed" can be interpreted either of "depending on the network need, it can be sent" or "the need is not concluded in RAN1". Just to refer "See Note 6" would be better.For CSI prediction- In LCM purpose column, "Inference" is started from capital but "tranining" and "monitoring is not from capital. Alighment among them is necessary like "Training" and "Monitoring". - To avoid to use the term "if needed" as it can be interpreted as "depending on the network need, it can be used". For beam management- In LCM purpose column, "Inference" is started from capital but "tranining" and "monitoring is not from capital. Alighment among them is necessary like "Training" and "Monitoring". - To avoid to use the term "if needed" as it can be interpreted as "depending on the network need, it can be used". - Note 3 looks FL comment description. Probably following is more reasonable?Note 3: ~~Please carefully n~~Note the ~~usage of “from Set A” vs. “for set B” in the table. T~~he usage of “from Set A” does not mean ~~reflect the fact that not~~ all Set A beams but can be ~~are needed and~~ a subset of beams ~~from Set A may be enough~~.For positining- In LCM purpose column, "Inference" is started from capital but "model tranining" and "monitoring is not from capital. Alighment among them is necessary like "Model training" and "Monitoring". - "Feasibility and necessity are under discussion" in monitoring column is covered already by Note 8?- " Example of calculation on the lower bound per PRS/SRS ....", " Example of calculation of the upper bound per PRS/SRS ....:", and so on are under Note 3. Current text style of under bar is same as higher level text style of "for positioning" and so on. Something different style should be used and these description shoudl be indended to clarify it is under note 3.- "If needed" usage in Note 8 should be avoided as it can be interpreted as "depending on the network need, it can be used". For "Common Notes for all sub-use-cases"- In the bullet of "in RAN1's answer... ", "R1-2308730" is used. In the bullet of "in this reply .. ", " RAN1 response to part A" is used. Both are same. To use "RAN1 response to part A" would be more consistent and friendly to RAN2 (as they don't care RAN1 tdoc number).- The bullet of "in addition, there ..." can be described as following. Or if these "in addition" column is intended to be under "in RAN1's answer" bullet, these should be indended further.* ~~In addition, there may be other information useful f~~For training at the UE-side, NW side, or neutral-side, there may be other information useful such as timestamp, data quality, UE speed, SNR, and vendor-specific information but ~~. Such information is~~ not included in the ~~following~~ tables, as RAN1 is still discussing whether their standardization is required ~~For~~ One of example is ~~, in~~ positioning enhancement~~., some information has been considered as potential spec impact in agreement of RAN1-114 meeting~~ (e.g., quality indicators, time stamps, [scenario identifier, LOS/NLOS condition, timing error])~~.~~
* ~~In addition, t~~There may be configuration information that may be collected together with data. Such information is not included in the tables. Examples include RS configuration(s) for deriving measurements, configuration related to Set A and/or Set B, information on association/mapping of Set A and Set B.
*

- I think rewording like below would be more understanable to RAN2. * For monitoring, RAN1 provided replies only for near-real-time monitoring ~~only~~. ~~RAN2 can consider t~~The requirements for data collection for relaxed monitoring can be considerd to be similar to offline training requirements.

- Following bullet can be coverd by adding "It reflects the current status of RAN1 discussion" in the beginning of LS reply simialr to "RAN1 response to part A"?- This LS reply is meant to capture existing RAN1 agreements/conclusions/observations and discussions for the purpose of replying the RAN2 LS; The LS reply does not serve as additional agreements/conclusions/observations beyond what RAN1 has already agreed/concluded/observed. |
| NTT DOCOMO | For the alignment of CSI compression and CSI prediction, Note 4 for CSI prediction should be updated into one for CSI compression as follows, which reflects the current RAN1 study status more precisely.Note 5: There is no agreement on the exact metric, monitoring output, and reporting format. An example based on companies’ evaluations is: SGCS (10s of bits)~~Note 5: There is no agreement on the performance metric or monitoring output details.~~ |
| Lenovo | For the training data of CSI compression, we suggest to remove the “first” or “second” stage of Type-3. As the discussed Type-3 is only one example.If the group prefer to keep it for “CSI feedback” for clarification we can have it as one example.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LCM purpose** | **Data content** | **Typical data size (per data sample)** | **Typical latency requirement** | **Notes** |
| training | Target CSI  | See Notes 1, 2 | Relaxed | This row applies to Type 1, Type 2, and ~~the first or second stage of~~ Type 3 separate training. |
| CSI Feedback | See Note 3 | Relaxed | This is for dataset delivery ~~for the second stage of~~ Type 3 (e.g., for the second stage of example implementation of Type-3) separate training and forward propagation information for Type 2 training. |

  |
| Huawei, HiSilicon | **For CSI compression:**Comments:1) For Type 2 related descriptions, as we already have conclusion to deprioritize Type 2 over the air interface, which means there is no spec impact in RAN (both RAN1/2), we do not need to mention it in the table. 2) For Type 3 description, it seems only NW first training is involved (target CSI + CSI feedback as the 2nd stage). For UE first training where Reconstructed CSI + CSI feedback as the 2nd stage, Reconstructed CSI is missed from the training entry.3) For Note 2, the example is changed with the majority evaluation assumption (also see referred table in CSI eva observations in TR): rank 1, 13 subbands. The numbers of CSI payloads are also adjusted.4) For Note 3, CSI feedback payload size is changed to “up to ~1000” – note that the small CSI payload in legacy can be <100bits.5) For Monitoring, reconstructed CSI from NW to UE, the “expected to be similar to target CSI for monitoring” is moved to [ ], since there is ongoing discussion at 8.14.3; if it is precoded on CSI-RS, then there is no need to discuss the size at RAN2.Changes:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LCM purpose** | **Data content** | **Typical data size (per data sample)** | **Typical latency requirement** | **Notes** |
| training | Target CSI  | See Notes 1, 2 | Relaxed | This row applies to Type 1, Type 2, and the first or second stage of Type 3 separate training. |
| CSI Feedback | See Note 3 | Relaxed | This is for dataset delivery for the second stage of Type 3 separate training ~~and forward propagation information for Type 2 training~~. |
| ~~Gradients for CSI Feeback~~ | ~~No agreement~~ | ~~Relaxed~~ | ~~This is for backward propagation for Type 2 training~~ |
| Inference | CSI Feedback | See Note 3 | Time-critical | Can use L1 report similar to legacy CSI |
| monitoring | Reconstructed CSI from NW to UE (if needed) | No agreement; [expected to be similar to target CSI for monitoring] | Near-real-time | This is called “UE-sided monitoring” in RAN1. |
| Calculated performance metrics (if needed) | See Note 4 | Near-real-time | This is called “UE-sided monitoring” in RAN1. |
| Target CSI (if needed) |  See Note 2 | Near-real-time | This is called “NW-sided monitoring” in RAN1. |

Note 1: Target CSI may be precoding matrix or channel matrix. RAN1’s reply for data size is based on precoding Matrix which has been more widely evaluated than channel matrix.Note 2: Data size for target CSI depends on the format. There is no agreement on the format of the target CSI. Some examples based on companies’ evaluations are: eType-II format (~1000 bits), eType-II-like format (~ a few 1000 bits), and float32 format (~ ~~150K~~ 50K bits). The data size may also vary depending on the scenario / configuration, and the captured value indicates the order of magnitude of the typical data size per sample as a guideline. As examples:* In eType-II PC 8 format, the payload size (PMI part) for rank 1, 13 subbands, 32 ports is around 300bits ~~rank 2, 19 subbands, 32 ports is around 800 bits~~.
* In floating point format (32 bits per sample), the precoding matrix for rank 1, 13 subbands, 32 ports is around 50K bits ~~4 layers, 19 subbands, 32 ports needs around 150 kilobits~~. This number doesn’t account for any potential compression techniques.

Note 3: There is no agreement on the CSI feedback size. Values in the order of eType II payload size may be assumed (up to ~ 1000 bits) for RAN2 discussion.Note 4: There is no agreement on the exact metric or reporting format. An example based on companies’ evaluations is: SGCS (10s of bits)Note 5: There are no agreements on the reporting type.Note 6: Feasibility and necessity of the monitoring schemes listed in the table are under discussion**For CSI prediction**Comments:1) For Note 2, the example is changed to 13 subbands for alignment with CSI compression. The numbers of CSI payloads are also adjusted.2) For Note 3, CSI feedback payload size is changed to “up to ~1000” – note that the small CSI payload in legacy can be <100bits.Changes: Note 1: Target CSI may be precoding matrix or channel matrix. RAN1’s reply for data size is based on channel matrix which has been more widely evaluated than precoding Matrix.Note 2: Data size for target CSI depends on the format. There is no agreement on the format of the target CSI. The data size may also vary depending on the scenario / configuration, and the captured value indicates the order of magnitude of the typical data size per sample as a guideline. As examples:* In floating point format (32 bits per sample), the channel matrix for 4 layers, 13 subbands ~~19 subbands~~ (one matrix per subband), 32 ports needs around 100 ~~150~~ kilobits per CSI-RS instance. Assuming 10 CSI-RS observation instances as input to predict one future CSI instance, the total is around 1M ~~1.5M~~ bits. This number doesn’t account for any potential compression techniques.

Note 3: There is no agreement on the predicted CSI feedback size. Values in the order of eType II payload size may be assumed (up to ~ 1000 bits) for RAN2 discussion.Note 4: There are no agreements on the reporting type.Note 5: There is no agreement on the performance metric or monitoring output details.Note 6: Feasibility and necessity of the monitoring schemes listed in the table are under discussion.**For beam management**Comments:For Note 1, there is no example for Beam ID?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LCM purpose** | **UE-side/NW-side models** | **Data content** | **Typical data size (per sample)** | **Typical latency requirement** | **Notes** |
| training | UE-side, NW-side | L1-RSRPs [and beam-IDs] for Set BL1-RSRPs and/or beam-IDs from Set A | See Note 1 for L1-RSRPs | Relaxed |  |

Note 1: There is no agreement on the data size of L1-RSRPs for Set A or Set B, but the following typical data size is provided as guidance for RAN2 discussion. Based on existing L1-RSRP reporting methodology, i.e., 7 bits for the strongest beam and 4 bits for the remaining beams, for Set B = 16 as an example, the typical data size would be 67 (hence up to ~100 bits), and for Set A = 128 as an example, the typical data size would be 515 (hence up to ~500 bits). For BM Case 2, the data size L1-RSRPs for Set A represents the data size per predicted future time instance. Note 2: There are no agreements on the reporting type.Note 3: Please carefully note the usage of “from Set A” vs. “for set B” in the table. The usage of “from Set A” reflect the fact that not all Set A beams are needed and a subset of beams from Set A may be enough.**For positioning**Comments:1) For Note 2/Note3, it is not clear which “appropriate WG” is to further discuss the dimension of the data types. Make it more generic.2) For Note 8, the #114 observation on monitoring data/metrics have a note “Note2: there’s no consensus during SI on whether monitoring metric will have spec impact or not”. This is reflected to Note 8, i.e. the further discussion includes not only the feasibility, but the need of spec impact. In addition, it says “RAN1 will continue discussing” but in the end says “in appropriate WG”, which are contradictory. Thus, “in appropriate WG” is removed.Changes: Note 2: For measurements as model input, no agreement on measurement types (i.e., time, power, and/or phase) in RAN1 for all cases (i.e., Case1 to Case3b). Measurement types (including their necessity) and sizes/dimension needs to be further discussed ~~in an appropriate working group~~. Candidate measurement types discussed/evaluated for model input include CIR (contains timing, power and phase information), PDP (contains timing and power information), DP (contains timing information). For labels (i.e., model output) of AI/ML assisted positioning (Case2a, Case3a), RAN1 identified an initial listing of candidates that provide performance benefits (i.e., timing info, [RSRP/RSRPP], LOS/NLOS indicator).Note 3: The measurement size of one data sample = (measurement data size of one PRS/SRS resource)\*(number of PRS/SRS resources needed for model input). The quantization and bit representation of time, power, and phase information (including their necessity) still need to be further discussed ~~in an appropriate working group~~. As a reference to existing timing and power representation, multipath measurement reporting for measurements of one PRS resource in the existing specification TS 37.355 is shown below for UE reporting to LMF. Similar measurement reporting exists in TS 38.455 for gNB reporting to LMF for one SRS resource:nr-DL-PRS-RSRP-Result-r16  7 bitsnr-DL-PRS-RSRP-ResultDiff-r16  6 bitsnr-RSTD-r16  16 to 21 bitsnr-RelativeTimeDifference-r16  9 to 14 bits……Note 8: RAN1 will continue discussing further details on monitoring data/metrics (including their feasibility and whether spec impact is needed) and mapping to different AI/ML positioning cases (if needed) ~~in appropriate working group~~. No agreement on the need or content of data collection for the purpose of model monitoring. This is to be discussed separately for Cases 1/2a/2b/3a/3b. These are descriptions on metrics that have been studied by RAN1:  - Statistics of the difference between model output and provided (approximate) ground truth label,  - Statistics of measurement and/or model input compared to the statistics associated with the training data - Statistics of model output compared to the statistics associated with the training data and/or its own previous inference output.**Common Notes for all sub-use-cases:**Comments:As we already provide examples in later part, there is no need to add the information in the first sentence. Note that “UE speed”, “vendor specific information” are not present in RAN1 agreements.Changes:* In addition, there may be other information useful for training at the UE-side, NW side, or neutral-side, ~~such as timestamp, data quality, UE speed, SNR, and vendor-specific information~~. Such information is not included in the tables, as RAN1 is still discussing whether their standardization is required. For example, in positioning enhancement, some information has been considered as potential spec impact in agreement of RAN1-114 meeting (e.g., quality indicators, time stamps, [scenario identifier, LOS/NLOS condition, timing error]). As another example, for beam prediction, additional information for the contents of collected data was considered (e.g., timestamps, SNR, data quality, etc.).
 |
| Ericsson | Comments below are for the positioning use case.**The following need to be entered in the column “Typical data size (per data sample)”.** Putting them in Notes is inappropriate since the size of one training data sample is not provided in the table. Also the yellow-highlighted bullet below is missing (i.e., not in Notes either) for AI/ML assisted positioning (Case 2a,3a) for training and inference.Size of one training data sample = (measurement data size of one PRS/SRS resource)\*(number of PRS/SRS resources needed for model input)Size of one training data sample = (label data size of one PRS/SRS resource)\*(number of PRS/SRS resources needed for model output)For Note 3:It’s not clear why “power/real value info”. Power is never represented by complex value. Change to either “power (real value)” or simply “power”.“where N is number of PRS/SRS resources ~~to consider~~ used as model input for obtaining a positioning fix.”For Example of calculation of the upper bound per PRS/SRS resource: No company ever suggested or evaluated CIR input with 256 samples that are not consecutive in a time window. If using the max of 256 samples as model input, they are always consecutive in time and spaced according to the sampling period. Thus, there is no need to send 256 individual timing values (i.e., delete (21 + 14\*255)). Also, it is necessary to explain the assumption used in the calculation, since there is no existing spec on how to send CIR values (complex values). Thus, **change to the following**:A potential upper bound on measurement size per PRS/SRS resource can be calculated as follows (assuming timing, power, and phase for 256 consecutive samples per PRS/SRS resource): 2\*(7 + 6\*255) = 3074 bits. Here it is assumed that the two values of one CIR complex-valued sample (either (magnitude, phase) or (real, imaginary)) are each represented using the same number of bits as the power value (real) in the existing specification 37.355. The total upper bound can be 3074\*N bits, where N is number of PRS/SRS resources used as model input for obtaining a positioning fix.For location information coordinates (corresponding to model output)label: do not under why mention “(e.g., velocity, location error, integrity info, etc.)”. On the other hand, it is necessary to mention other info in existing LocationInformation IE since this IE is likely to be reused. Add:It is noted that the existing LocationInformation IE in 37.355 include more elements beyong locationCoordinates, including: measurementReferenceTime, locationSource.For intermediate positioning measurement (corresponding to model output): change from 28 bits to 21 bits, since majority companies used only timing info as mode output, (not timing info + power info). In fact power info [RSRP/RSRPP] is put in brackets in text. Also, for assisted positioning, N is not necessarily the number for obtaining a positioning fix. For example, there can be 18 models for a positioning fix, each model covers one TRP (each model has 1 intermediate measurement as model output)Thus, **change to the following**:As a reference to existing timing representation in Rel17 [TS 37.355], an example on the label size can be of 21 bits per PRS/SRS resource while assuming model output produces one timing of 21 bits per PRS/SRS resource. The label size can be 21\*N bits, where N is number of PRS/SRS resources for which intermediate positioning measurement has been generated.For monitoring, data content, It is incorrect to imply that monitoring metrics is the data content for monitoring. **Change to the following**: “~~RAN1 has studied initial listing of monitoring metrics~~ No agreement on the need or content of data collection for the purpose of model monitoring. This is to be discussed separately for Cases 1/2a/2b/3a/3b.“ **For Note 6, it is incorrect to say that such list has been identified by RAN1.** It is stated clearly in RAN1 agreement below that LMF is identified only as candidates when monitoring is based on provided ground truth label (or its approximation).Agreement (RAN1#112bis)Regarding monitoring for AI/ML based positioning, at least the following entities are identified to derive monitoring metric* UE at least for Case 1 and 2a (with UE-side model)
* gNB at least for Case 3a (with gNB-side model)
* LMF at least for Case 2b and 3b (with LMF-side model)

AgreementRegarding AI/ML model monitoring for AI/ML based positioning, the following entities are identified as candidates to derive monitoring metric in addition to entities from previous agreement* LMF for Case 2a (with UE-side model) and Case 3a (with gNB-side model) at least when monitoring is based on provided ground truth label (or its approximation)

**Delete Note 8.** It is incorrect to imply that monitoring metrics is the data being collected for monitoring. Also, it is incorrect to say that RAN1 will continue discussing, since “Other aspects of positioning” is already closed for the Rel-18 study item. |
| Mod | To NTT DOCOMO: There were no such evaluations for CSI prediction.To Huawei:1) For Type 2 related descriptions, as we already have conclusion to deprioritize Type 2 over the air interface, which means there is no spec impact in RAN (both RAN1/2), we do not need to mention it in the table. 🡺 As stated earlier, we are not making judgements on spec impact in other working groups. In this particular case, potential spec impact to SA has been mentioned by companies.I didn’t find time to address your comments on the data size. Let’s discuss later.To Ericsson:For monitoring, data content, It is incorrect to imply that monitoring metrics is the data content for monitoring. **Change to the following**: “~~RAN1 has studied initial listing of monitoring metrics~~ No agreement on the need or content of data collection for the purpose of model monitoring. This is to be discussed separately for Cases 1/2a/2b/3a/3b.“ 🡺 Please note that we are including monitoring metrics in the table in all the other sub-use-cases.I didn’t find time to address your comments on the data size. Let’s discuss later. |
| Lenovo | In fact, separate training (i.e., Type-3) is not limited to the “example Type-3” that we have evaluated for CSI-feedback compression. There are other types of Separate training as well. It was the reason than in the previous round of comments we have requested the following changes. We are okay to keep it as an example though.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LCM purpose** | **Data content** | **Typical data size (per data sample)** | **Typical latency requirement** | **Notes** |
| training | Target CSI  | See Notes 1, 2 | Relaxed | This row applies to Type 1, Type 2, and ~~the first or second stage of~~ Type 3 separate training. |
| CSI Feedback | See Note 3 | Relaxed | This is for dataset delivery ~~for the second stage of~~ Type 3 (e.g., for the second stage of example implementation of Type-3) separate training and forward propagation information for Type 2 training. |

 |
| Samsung | **Some comments for beam management**1. data content for training, we don’t agree. We’d like to provide the reason again: There are different labelling options for data training, current answer will miss leading RAN 2 that “L1-RSRPs of All Set is needed”. Even in NOTE 3, the usage of “from Set A” reflect the fact that not all Set A beams are needed and a subset of beams from Set A may be enough. We suggest to change it to

L1-RSRPs and~~/or~~ [beam-IDs] from Set A or Top-1 or Top-K beam-ID(s) from Set AAlternatively, we suggest to paste the agreements in note 3.

|  |
| --- |
| - Option 1a: Top-1 beam(pair) in Set A and L1-RSRPs of Set B with implicit or explicit beam IDs- Option 1b: Top-K beam (pair)s in Set A and L1-RSRPs of Set B with implicit or explicit beam IDs - Option 2a: L1-RSRPs per beam of all the beams(pairs) in Set A - Option 2b: Top-K beam(pair)s in Set A and the corresponding L1-RSRPs and L1-RSRPs of Set B with implicit or explicit beam IDs - Option 2c: Top-1 beam(pair) in Set A and the corresponding L1-RSRP and L1-RSRP of Set B with implicit or explicit beam IDs |

1. For data content for inference for UE-side, there is no discussion/agreement on “Confidence/probability information related to predicted beams (if supported)”. Instead, we think this should belong to “monitoring”
2. For latency requirements for monitoring, we think it could be “Time-critical”, e.g., for “Confidence/probability information related to predicted beams” as metric for LCM.
3. What is “Event occurrence and/or calculated performance metrics (if needed)” for monitoring for UE side?
4. For monitoring, the description of “…from Set A”. We suggest to provide detailed examples, and this could be same or different from data collection, which should also be clarified.
 |
| Xiaomi | **CSI prediction**For CSI prediction, Note 3 only provides the data size of predicted CSI. But it does not include the data size of the corresponding ground truth. We think other note should be added to clarify, i.e.,Note 7: Data size for ground truth depends on the format. There is no agreement on the format of the ground truth.**Beam management**For monitoring, as for the 2nd row, UE need to report the data based on measurement and AI/ML model output for performance metric calculation at NW-side. It means that the data size will be the double of L1-RSRPs and/or beam-IDs of beams from Set A. While for the 3rd row, UE need to report both set B for model input and set A based on measurement.For note 1, similar description on set B is also needed.For BM case 2, in order to provide more information to RAN2, it is better to provide some typical value of the number of future prediction time instance and history measurement time instance.Thus, we suggest the following update:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LCM purpose** | **UE-side/NW-side models** | **Data content** | **Typical data size (per sample)** | **Typical latency requirement** | **Notes** |
| Training | UE-side, NW-side | L1-RSRPs [and beam-IDs] for Set BL1-RSRPs and/or beam-IDs from Set A | See Note 1 for L1-RSRPs | Relaxed |  |
| Inference | UE-side | Predicted L1-RSRPs (if supported) and/or predicted beam-IDs from Set AConfidence/probability information related to predicted beams (if supported) | Small (10s of bits) | Time-critical | RAN1 has agreed to consider L1 signalling for this reporting |
| NW-side | L1-RSRP, [Beam-ID] for Set B | See Note 1 for L1-RSRPs | Time-critical |
| Monitoring | UE-side | Event occurrence and/or calculated performance metricsSee Note 4 | Small (10s of bits) | Near-real-time |  |
| UE-side | L1-RSRPs and/or beam-IDs of beams from Set A based on measurement and AI/ML model outputSee Note 4 | See Note 1 for L1-RSRPs | Near-real-time | FFS: The data based on measurement and the data from model output will be reported in one report or not. |
| NW-side  | L1-RSRP, [Beam-ID] for Set B L1-RSRPs and/or beam-IDs of beams from Set ASee Note 4 | See Note 1 for L1-RSRPs | Near-real-time | FFS: The data for set B and the data from set A will be reported in one report or not. |

Note 1: There is no agreement on the data size of L1-RSRPs for Set A or Set B, but the following typical data size is provided as guidance for RAN2 discussion. Based on existing L1-RSRP reporting methodology, i.e., 7 bits for the strongest beam and 4 bits for the remaining beams, for Set B = 16 as an example, the typical data size would be 67 (hence up to ~100 bits), and for Set A = 128 as an example, the typical data size would be 515 (hence up to ~500 bits). For BM Case 2, the data size of L1-RSRPs for Set A represents the data size per predicted future time instance. And the data size of L1-RSRPs for set B represents the data size per history measurement time instance.Note 2: There are no agreements on the reporting type.Note 3: The usage of “**from** Set A” reflect the fact that not all Set A beams are needed and a subset of beams from Set A may be enough. The usage of “**for** Set B” reflect the fact that all Set B beams are needed.Note 4: Feasibility and necessity of the monitoring schemes listed in the table are under discussion.Note 5: For BM Case 2, the typical value of the number of history measurement time instance is [2~8] and typical value of the number of predicted future time instance is [1~4].**Positioning**For Note 7, we think the term of “monitoring entity” is not clear. Does that mean the entity collecting data for monitoring metric calculation or the entity calculate the monitoring metric or both? **Common Notes for all sub-use-cases**For the notes of monitoring, it provides one hint that relaxed monitoring will be supported as well. That would be misleading. The following is our suggestion * For monitoring, RAN1 provided replies only for near-real-time monitoring. The requirements for data collection for relaxed monitoring(if feasible and necessary) can be considered to be similar to offline training requirements.
 |
| ZTE | We share similar view as Huawei’s 1st comment regarding Type 2 related descriptions. Our preference is to remove the Type 2 related descriptions. But if companies prefer to keep them, another way is to add a note to clarify that RAN1 has agreed to deprioritize Type 2 over the air interface. |
| OPPO | For CSI compression1. It is better is to make it clear that “CSI feedback” is from UE to NW, or NW to UE, or both. Thus, the following wording change is suggested:

This is for dataset delivery from NW to UE and from UE to NW for the second stage of NW-first and UE-first Type 3 separate training respectively and forward propagation information for Type 2 training.For CSI prediction1. It is better to make it more accurate that the discussion is only for UE-side model. Thus, the following wording change is suggested:

For CSI prediction with UE-side modelFor beam management1. UE may report all Set A beams or part of Set A beams. It is up to the final design. The current Note 3 says only reporting of part of Set A beams is supported, which is incorrect. Thus, the following wording change is suggested:

Note 3: The usage of “**from** Set A” can include ~~reflect~~ the case ~~fact~~ that not all Set A beams are needed and a subset of beams from Set A may be enough. The usage of “**for** Set B” reflect the fact that all Set B beams are needed. |
| CATT | Generally fine with the current version.1) For CSI prediction, since we delete the ‘generation entity’ and there is no column for ‘model location’ in the table, we agree with OPPO’s version of ‘CSI prediction with UE-side model’ at the title.2) For Positioning, for monitoring, Note 8 should also be quoted in the last column: See Note 6, 7 and 8 |
| Mod | To Lenono: We can focus on typical training types discussed in RAN1. To Samsung: 1. I think “L1-RSRPs and/or beam-IDs from Set A” is more generic and covers all the options.2. There is an agreement in 9.2.3.23. For all the monitoring, we put near-real-time. If we put time-critical, it would mean that the monitoring is at the same frequency as inference, which does not seem to make sense.4. Agreement from RAN1 #113:  AgreementFor BM-Case1 and BM-Case2 with a UE-side AI/ML model, regarding performance monitoring, study potential spec impact(s) from the following aspects in addition to those included in previous agreements:* Configuration/Signalling from gNB to UE for measurement and/or reporting
* UE calculates performance metric(s), either reports it to NW or reports an event to NW based on the performance metric(s)
	+ FFS: definition of an event and the performance metric(s) used to identify it
* Indication from NW for UE to do LCM operations

5. Do you see the need of going into further details? The data size is only up to ~500 bits. What different would further details make to RAN2?To Xiaomi: For monitoring, the convention we’re using is that we don’t list data that’s already provided via inference. |
| AT&T | We can align the CSI prediction and CSI compression note regarding the Calculated performance metrics. The companies provided the SGCS and NMSE as intermediate KPI for their evaluation. This metric can be computed directly at the UE (without NW involvement) and can be used as performance metric for the model. It is intended to provide clarity to RAN2 regarding size of the calculated performance metrics. The note 5 for CSI prediction should be updated to Note 5: There is no agreement on the exact metric or reporting format. An example based on companies’ evaluations is: SGCS (10s of bits)  |
| Lenovo | In fact, separate training (i.e., Type-3) is not limited to the “example Type-3” that we have evaluated for CSI-feedback compression. There are other types of Separate training as well. It was the reason than in the previous round of comments we have requested the following changes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LCM purpose** | **Data content** | **Typical data size (per data sample)** | **Typical latency requirement** | **Notes** |
| training | Target CSI  | See Notes 1, 2 | Relaxed | This row applies to Type 1, Type 2, and ~~the first or second stage of~~ Type 3 separate training. |
| CSI Feedback | See Note 3 | Relaxed | This is for dataset delivery ~~for the second stage of~~ Type 3 separate training and forward propagation information for Type 2 training. |

If you want to elaborate more, we are okay to keep it as an example as below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LCM purpose** | **Data content** | **Typical data size (per data sample)** | **Typical latency requirement** | **Notes** |
| training | Target CSI  | See Notes 1, 2 | Relaxed | This row applies to Type 1, Type 2, and ~~the first or second stage of~~ Type 3 separate training. |
| CSI Feedback | See Note 3 | Relaxed | This is for dataset delivery ~~for the second stage of~~ Type 3 (e.g., for the second stage of example procedure for Type-3 training explained in Section 6.2.1 of the TR 38.843 v1.0) separate training and forward propagation information for Type 2 training. |

 |
| Intel | For CSI compression use-case we think example with float32 is an extreme case, suggest to change it to 8-bit scalar (real value) quantization, see below from R1-2306832 |
| Xiaomi | **CSI feedback**In principle, we are fine with the current table. But the following comments are provided for further clarification. * **CSI compression**

Regarding CSI compression monitoring, the data size of target CSI should see Note 1 and Note 2. In current table, it only sees Note2.* **CSI prediction**

Similarly with above comment, the data size of target CSI for CSI preidiction monitoring should also see Note 1 and Note 2. Note 2 is missed in current table.In addition, for the Note 3 of CSI prediction, the values in the order of eType II payload size should be assumed (up to ~ 1000 bits) for RAN2 discussion.**Positioning**For Note 7 in positioning, we don’t think the statement is accurate. At least agreement listed in Note 6 give the information about the entity for the monitoring metric calculation entity (Assume monitoring entity includes the entity calculating the monitoring metric in the note) and potential mapping between monitoring metric calculation entity and inference entity (i.e., at least inference entity can perform the monitoring metric calculation). So the following is our suggestion for update. Note 7: No agreement yet ~~on a monitoring entity and/or~~ a monitoring decision entity or their mapping to other entities (e.g., entity running the inference, entity deriving the monitoring metric, etc.).**Common Notes for all sub-use-cases**For the monitoring, we could understand the motivation of near-real-time monitoring but we doubt the feasibility and necessity of relaxed monitoring and there is no agreement or conclusion to support relaxed monitoring. So that’s why we suggest the following update. * For monitoring, RAN1 provided replies only for near-real-time monitoring. If relaxed monitoring is feasible and necessary, The requirements for data collection for relaxed monitoringcan be considered to be similar to offline training requirements.
 |
| ~~New H3C~~ | For CSI compression ,we can’t find “note 5” in the table so we suggest removing note 5~~Note 5: There are no agreements on the reporting type.~~ |
| Samsung | For beam management

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LCM purpose** | **UE-side/NW-side models** | **Data content** | **Typical data size (per data sample)** | **Typical latency requirement** | **Notes** |
| Training | UE-side, NW-side | L1-RSRPs [and beam-IDs] for Set BL1-RSRPs and/or beam-IDs of a subset (including one beam) or full set of Set A | See Note 1 for L1-RSRPs | Relaxed |  |
| Inference | UE-side | Beam prediction results | Small (10s of bits) | Time-critical | RAN1 has agreed to consider L1 signalling for this reporting |
| NW-side | L1-RSRP, [Beam-ID] for Set B | See Note 1 for L1-RSRPs | Time-critical |
| Monitoring | UE-side | Event occurrence and/or calculated performance metrics See Note 4 | Small (10s of bits) | Near-real-time |  |
| UE-side | L1-RSRPs and/or beam-IDs of beams from a subset of Set A (including one beam) See Note 4 | See Note 1 for L1-RSRPs | Near-real-time |  |
| NW-side  | L1-RSRPs and/or beam-IDs of beams from a subset of Set A (including one beam)See Note 4 | See Note 1 for L1-RSRPs | Near-real-time |  |

Note 1: There is no agreement on the data size of L1-RSRPs for Set A or Set B, but the following typical data size is provided as guidance for RAN2 discussion. Based on existing L1-RSRP reporting methodology, i.e., 7 bits for the strongest beam and 4 bits for the remaining beams, for 16 beams as an example, the typical data size would be 67 (hence up to ~100 bits), and for 128 beams as an example, the typical data size would be 515 (hence up to ~500 bits). For BM Case 2, the data size L1-RSRPs for Set A and Set B represents the data size per predicted future time instance and per history measurement time instance, respectively. Note 2: There are no agreements on the reporting type.Note 4: Feasibility and necessity of the monitoring schemes listed in the table are under discussion.Note 5: For BM Case 2, the typical value of the number of history measurement time instance used in evaluations is [1~8] and typical value of the number of predicted future time instance is [1~4]. |
| **Mod** | **To SS: I incorporated your changes, except the change on Note 5. I don’t think the history length 1 makes sense, as the AI/ML will become the same as sample-and-hold.** |