3GPP TSG-RAN WG2 #122 R2-23XXXXX

Incheon, Republic of Korea, May 22 – 26, 2023

Agenda Item: 7.16.1

Source: Rapporteur

Title: Progress and Next Steps: Rapporteur's Insights

Document for: Information

# 1 Introduction

The RAN1-led Study Item (SI) on *“AI/ML for NR Air Interface”* was approved in RAN#94-e in December 2021 (see the approved SID in [RP-213599](http://www.3gpp.org/ftp//tsg_ran/TSG_RAN/TSGR_94e/Docs//RP-213599.zip) and the revised version in [RP-221348](http://www.3gpp.org/ftp//tsg_ran/TSG_RAN/TSGR_96/Docs//RP-221348.zip)).

According to the time budget presented in the Table below (see [RP-221060](http://www.3gpp.org/ftp//tsg_ran/TSG_RAN/TSGR_96/Docs//RP-221060.zip) for further details), RAN2 is currently at the halfway point of the SI. Therefore, the rapporteur believes that this document could serve as a means to reflect on the SI's progress and provide some ideas on how the discussion could proceed.



The SI focuses on analyzing enhancements to the following use cases:

* CSI feedback,
* beam management, and
* positioning accuracy.

As understood from the SID and RAN1’s Work Plan (in [R1-2205021](http://www.3gpp.org/ftp//tsg_ran/WG1_RL1/TSGR1_109-e/Docs//R1-2205021.zip)), RAN1’s focus is on:

* Evaluating the performance benefits of using AI/ML over existing frameworks,
* Describe AI/ML framework(s), model(s) and terminologies,
* Assess specification impact.

While from the SID, it is understood that RAN2’s work centers around assessing potential specification impact towards protocol aspects while studying solutions to address such impacts.

RAN2’s discussions have been organized as follows:

1. AIML methods, and
2. Use case specific aspects.

It should be noted that although there has been a strong willingness to discuss use case-specific aspects, the clear dependency on RAN1's evaluations and related requirements has prevented RAN2 from adequately addressing the point mentioned above. As a result, only a few agreements and assumptions have been recorded in the meeting notes.

Thus far RAN2’s discussions have primarily focused on AIML methods, specifically on the following:

* AIML model-specific matters, including:
	+ Identification (e.g., model ID),
	+ Meta information,
	+ Format,
* Data collection, which includes
	+ Identifying and analyzing existing frameworks
		- Note: special attention has been given to model training and monitoring, while "inference" appears to be becoming relevant to the discussion as well.
* Model transfer/delivery, including:
	+ Exploring alternatives based on the use cases' architectures and scenarios,
	+ Functional block mapping

In addition, the latest RAN2 meeting (RAN2#121bis-e) has initiated efforts to establish a functional framework for Life Cycle Management (LCM) that can structure and organize the AIML methods listed above. This framework is intended to cover related procedures that have emerged during RAN2 discussions, such as UE capability reporting, model/functionality selection/(de)activation/switching/fallback, and so on.

The full list of agreements can be found in the Annex for completeness. Additionally, please refer to the References section of this document for the related Status Reports presented in RP which also include contributions from all participating companies in previous meetings.

# 2 Structuring the discussion

## 2.1 Progress and planning

Below you will find a draft of what the planning for future meetings could look like. However, since future discussions depend on the outcome of our discussion in the present meeting and on RAN1 progress, we would like then to particularly emphasize what the Rapporteur believe should be covered during RAN2#122.

### 2.1.1 Targeting RAN2#122

The Rapporteur believes that during RAN2#122 we should focus on the following.

#### Data collection

On the one hand, the discussion related to data collection during RAN2#121bis-e focused on further structuring the scope of the endorsed table where RAN2 will evaluate the different data collection methods (see [R2-2302286](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121/Docs//R2-2302286.zip)).

While on the other hand, RAN2 discussed on whether to consider the EVEX framework.

At this point, the Rapporteur believes that RAN2 already have a set of agreements that should allow for tangible and specific/technical “conclusions” to be reached during RAN2#122.

We expect then for companies to organize their contributions following the latest agreements (please refer to the Annex), so that RAN2 can analyze the limitations/benefits of the different frameworks.

Specifically, the Rapporteur believes that, after all, companies should focus on answering the following:

* Considering whether we have a NW-, UE- or two-sided model scenario:
	+ What are the (a priori / in principle) requirements expected to be met when collecting data for inference/monitoring/training?

Conducting this exercise would allow RAN2 to gain a comprehensive understanding of the expectations from a RAN2 standpoint before receiving specific requirements from RAN1. Furthermore, adopting this approach would allow RAN2 in determining whether a particular data collection framework could eventually be suitable for any of the listed LCM purposes in the concerning (sub)use case.

If companies prefer to improve readability by following the approach but do not want to incorporate their ideas into the table, a parallel discussion can be initiated to consolidate the common RAN2 understanding. This should eventually lead to endorsing a new version of the table including the three newly agreed columns and any additional modifications to the original one.

If time allows, it may be possible to capture RAN2’s discussion in an LS for RAN1.

1. The set of RAN2 agreements concerning data collection should allow for tangible and arguably more technical conclusions during RAN2#122.
2. Prioritize the data collection discussion.
3. For the data collection discussion and considering whether we have a NW-, UE- or two-sided model scenario, companies are requested to capture their general understanding of the expected data collection “requirements” for training/inference/monitoring in their contribution papers.

#### LCM mechanisms

To start with, the Rapporteur would like to point out the following agreement reached during the last RAN1 meeting:

|  |
| --- |
| Agreement* Study necessity, mechanisms, after functionality identification, for UE to report updates on applicable functionality(es) among [configured/identified] functionality(es), where the applicable functionalities may be a subset of all [configured/identified] functionalities.
* Study necessity, mechanisms, after model identification, for UE to report updates on applicable UE part/UE-side model(s), where the applicable models may be a subset of all identified models.
 |

The Rapporteur believes these are RAN2-centric mechanism. And in fact, we note that these issues have already been brought up in RAN2 (e.g., related to potential enhancements to UE capability reporting framework, or specifically related to mechanisms that could allow to indicate the applicability of AIML functionalities or models under changing conditions). Hence, we believe that it is within RAN2’s scope to study mechanism to report updates on the applicability of a functionality/model.

1. Gather companies’ views regarding mechanisms to report the applicability of AIML functionalities and/or models.

Furthermore, RAN1 is currently discussing two different LCM management approaches:

* Functionality-based LCM, and
* Model ID-based LCM.

For completeness, the concerning agreements from the previous RAN1 meeting (i.e., RAN1#112bis-e) are shown below (see [R1-2304168](http://www.3gpp.org/ftp//tsg_ran/WG1_RL1/TSGR1_112b-e/Docs//R1-2304168.zip)), where the Rapporteur has highlighted the aspects we think are RAN2-related:

|  |
| --- |
| Agreement* For AI/ML functionality identification and functionality-based LCM of UE-side models and/or UE-part of two-sided models:
	+ Functionality refers to an AI/ML-enabled Feature/FG enabled by configuration(s), where configuration(s) is(are) supported based on conditions indicated by UE capability.
	+ Correspondingly, functionality-based LCM operates based on, at least, one configuration of AI/ML-enabled Feature/FG or specific configurations of an AI/ML-enabled Feature/FG.
		- FFS: Signaling to support functionality-based LCM operations, e.g., to activate/deactivate/fallback/switch AI/ML functionalities
		- FFS: Whether/how to address additional conditions (e.g., scenarios, sites, and datasets) to aid UE-side transparent model operations (without model identification) at the Functionality level
		- FFS: Other aspects that may constitute Functionality
	+ FFS: which aspects should be specified as conditions of a Feature/FG available for functionality will be discussed in each sub-use-case agenda.
* For AI/ML model identification and model-ID-based LCM of UE-side models and/or UE-part of two-sided models:
	+ model-ID-based LCM operates based on identified models, where a model may be associated with specific configurations/conditions associated with UE capability of an AI/ML-enabled Feature/FG and additional conditions (e.g., scenarios, sites, and datasets) as determined/identified between UE-side and NW-side.
	+ FFS: Which aspects should be considered as additional conditions, and how to include them into model description information during model identification will be discussed in each sub-use-case agenda.
	+ FFS: Relationship between functionality and model, e.g., whether a model may be identified referring to functionality(s).
	+ FFS: relationship between functionality-based LCM and model-ID-based LCM
* Note: Applicability of functionality-based LCM and model-ID-based LCM is a separate discussion.
 |

Let us note that, while the above was not explicitly treated during the previous RAN2 meeting, the following RAN2#121bis-e agreements point towards that direction. So, the Rapporteur thinks that RAN2 could have a brief discussion about these approaches.

|  |
| --- |
| * For the CSI compression and beam management use cases, model/function selection/(de)activation/switching/fallback can be UE-initiated or gNB-initiated. FFS how the different cases are different (e.g. applicability to UE-sided vs network sided model).
* For the positioning use case, model/function selection/(de)activation/switching/fallback can be UE-initiated or LMF-/ gNB-initiated. FFS how the different cases are different (e.g. applicability to UE-sided vs network sided model).
 |

1. Have a brief discussion to clarify the RAN2-related aspects, concerning functionality-based LCM, and model ID-based LCM (e.g., mechanisms involved such as UE capability reporting, configuration, etc.).
2. RAN2 can now analyze signaling and protocols for functionality- and model ID-based LCM taking RAN2#121bis-e agreements as reference.

Regarding model IDs, and as seen in the Annex, this is a topic of great interest in RAN2. However, the Rapporteur wonders whether it would be straightforward to continue developing such topic fully without RAN1 making some progress on the subject. Note that this might even depend on other WGs, for which RAN2 could eventually need to trigger LSs. So, we believe that currently, the focus should perhaps be on other more pressing matters (see subsections below).

1. Regarding detailed model identification matters, we wonder whether further progress can be made without RAN1's (or eventually other WG’s) input, e.g., meta data, uniqueness of IDs, content/structure of IDs, etc.
2. Do not prioritize detailed model ID-specific aspects during RAN2#122.

#### Functional framework

The following was agreed during RAN2#121bis-e:

|  |
| --- |
| * The general AI/ML framework consist of, (i) Data Collection, (ii) Model Training, (iii) Model Management, (iv) Model Inference, and (v) Model Storage.

Chair: the following was almost agreed (leave it FFS for now): AI/ML functional architecture in Figure 1 in [R2-2303674](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121bis-e/Docs//R2-2303674.zip) is the baseline with the modification that Performance Monitoring is changed to Model Mgmt / Performance Monitoring. It is noted that the exact interactions may need some modification depending on how each piece of functionality is specified**.**  |

According to our understanding of the discussion and how it led to the agreement shown above, the Rapporteur realizes that the agreement simply aims to capture a description of general stages within the management of an AIML model/functionality. However, this does not represent in any way a fixed structure that must be followed when implementing these in the future.

In fact, it is possible to gather from the Chair's note (and from the discussion held during the meeting) that this idea of a fixed (and eventually standardized) structure is what prevented reaching an agreement regarding the functional framework for our study. Moreover, the Rapporteur would like to emphasize that it was not only the modification/merging of the Performance Monitoring and Model Management blocks that made some companies hesitate, but also the fact of having a potential functional block representing Model Storage.

Based on this, and to reduce conflicts, the Rapporteur suggests adopting a similar approach to the one taken by RAN3 during their study (see [TR 37.817](https://www.3gpp.org/ftp/Specs/archive/37_series/37.817/37817-h00.zip) clause 4.2) and only represent the blocks of data collection, training, and inference in the functional framework, as shown in the image below:



Figure 1. Functional Framework for RAN Intelligence (as taken from TR 37.817 clause 4.2)

Note that the above is a simplification of RAN2’s agreement. So, for the sake of e.g., model transfer/delivery discussions, companies should still understand that models might be located/stored in entities different from those performing inference, training, or the collection of data.

In fact, RAN2 could adopt similar “high-level principles” as those followed by RAN3 (see list below, taken from [TR 37.817](https://www.3gpp.org/ftp/Specs/archive/37_series/37.817/37817-h00.zip) clause 4.1). In this regard, the Rapporteur believes that by building, as far as possible, RAN2 discussion on such principles, we will be able to reduce the amount of conflict in our debates.

1. The functional framework is not a fixed structure and should allow for implementation-specific solutions when applying AIML functionalities in the future.
2. Companies should still understand that models may be in different locations from the entities performing inference, training, or data collection.
3. Adopt an approach similar to that of RAN3's study, where the blocks of data collection, training, and inference are presented in the functional framework.
4. Draw inspiration from the "high-level" principles of the RAN3 study (see TR 37.817). RAN2 could eventually capture something similar (tailored made for our SI) to reduce controversial and time-consuming discussions.

|  |
| --- |
| 4.1 High-level Principles The following high-level principles should be applied for AI-enabled RAN intelligence:- The detailed AI/ML algorithms and models for use cases are implementation specific and out of RAN3 scope.- The study focuses on AI/ML functionality and corresponding types of inputs/outputs. - The input/output and the location of the Model Training and Model Inference function should be studied case by case.- The study focuses on the analysis of data needed at the Model Training function from Data Collection, while the aspects of how the Model Training function uses inputs to train a model are out of RAN3 scope.- The study focuses on the analysis of data needed at the Model Inference function from Data Collection, while the aspects of how the Model Inference function uses inputs to derive outputs are out of RAN3 scope.- Where AI/ML functionality resides within the current RAN architecture, depends on deployment and on the specific use cases.- The Model Training and Model Inference functions should be able to request, if needed, specific information to be used to train or execute the AI/ML algorithm and to avoid reception of unnecessary information. The nature of such information depends on the use case and on the AI/ML algorithm. - The Model Inference function should signal the outputs of the model only to nodes that have explicitly requested them (e.g., via subscription), or nodes that take actions based on the output from Model Inference.- An AI/ML model used in a Model Inference function has to be initially trained, validated and tested by the Model Training function before deployment.- NG-RAN SA is prioritized; EN-DC and MR-DC are down-prioritized, but not precluded from Rel.18.- Functional framework and high-level procedures defined in this TR should not prevent from “thinking beyond” them during normative phase if a use case requires so.- User data privacy and anonymisation should be respected during AI/ML operation. |

Table 1. High-level principles from RAN3’s TR 37.817.

To wrap-up, it is worth noting that model transfer/delivery matters may be impacted by the results of the data collection and functional framework discussion. Therefore, the focus may be better placed on the other issues mentioned previously.

1. Model transfer/delivery discussion may depend on data collection and functional framework results, so priority should be given to other matters.

### 2.1.2 Timeline

#### RAN2#122 (1 TU)

* Prioritize the following topics:
	+ Data collection
		- Considering NW-, UE- or two-sided model scenarios; separately analyze the data collection requirements and solutions for the different LCM purposes
		- From the above, try to reach preliminary conclusions on suitability of the different identified data collection frameworks
	+ LCM mechanisms:
		- Reporting the applicability of AIML functionalities/models
		- Functionality- and model ID-based LCM
			* Signaling and protocols for each
	+ Functional framework
		- Discuss whether there is a need for one, if so, is RAN3’s approach acceptable?
* If time allows:
	+ Continue evaluating the pros and cons associated with the model delivery/transfer methods

#### RAN2#123 (1 TU)

* The following topics can be further discussed in RAN2#123 meeting
	+ Any non-resolved topics, e.g.:
		- If not clear yet, prioritize for a tangible outcome for the data collection discussion
			* E.g., recommended solutions
		- If sufficiently clear from the outcome of the data collection, LCM and functional framework discussions, try to conclude on model transfer/delivery
			* E.g., recommended solutions
		- On reporting the applicability of AIML functionalities/models
	+ UE capability information

#### RAN2#123bis-e (1 TU)

*Note: RAN1 should have finalized their study by this point. So RAN2 should be able to gather RAN1’s conclusions to eventually revisit RAN2’s agreements or potential gaps in our study.*

* The following topics can be further discussed in RAN2#123 meeting
	+ Any non-resolved topics
* Focus on TR drafting:
	+ Solutions
	+ Procedures

#### RAN2#124 (1 TU)

* Finalize TR
	+ Take on unresolved issues.
	+ Include recommendations towards normative work

# 3 Conclusion

In the previous sections we made the following observations:

[Observation 1 The set of RAN2 agreements concerning data collection should allow for tangible and arguably more technical conclusions during RAN2#122.](#_Toc134171638)

[Observation 2 Regarding detailed model identification matters, we wonder whether further progress can be made without RAN1's (or eventually other WG’s) input, e.g., meta data, uniqueness of IDs, content/structure of IDs, etc.](#_Toc134171639)

[Observation 3 The functional framework is not a fixed structure and should allow for implementation-specific solutions when applying AIML functionalities in the future.](#_Toc134171640)

[Observation 4 Companies should still understand that models may be in different locations from the entities performing inference, training, or data collection.](#_Toc134171641)

[Observation 5 Model transfer/delivery discussion may depend on data collection and functional framework results, so priority should be given to other matters.](#_Toc134171642)

Based on the discussion in the previous sections we propose the following:

[Proposal 1 Prioritize the data collection discussion.](#_Toc134171643)

[Proposal 2 For the data collection discussion and considering whether we have a NW-, UE- or two-sided model scenario, companies are requested to capture their general understanding of the expected data collection “requirements” for training/inference/monitoring in their contribution papers.](#_Toc134171644)

[Proposal 3 Gather companies’ views regarding mechanisms to report the applicability of AIML functionalities and/or models.](#_Toc134171645)

[Proposal 4 Have a brief discussion to clarify the RAN2-related aspects, concerning functionality-based LCM, and model ID-based LCM (e.g., mechanisms involved such as UE capability reporting, configuration, etc.).](#_Toc134171646)

[Proposal 5 RAN2 can now analyze signaling and protocols for functionality- and model ID-based LCM taking RAN2#121bis-e agreements as reference.](#_Toc134171647)

[Proposal 6 Do not prioritize detailed model ID-specific aspects during RAN2#122.](#_Toc134171648)

[Proposal 7 Adopt an approach similar to that of RAN3's study, where the blocks of data collection, training, and inference are presented in the functional framework.](#_Toc134171649)

[Proposal 8 Draw inspiration from the "high-level" principles of the RAN3 study (see TR 37.817). RAN2 could eventually capture something similar (tailored made for our SI) to reduce controversial and time-consuming discussions.](#_Toc134171650)

# 4 References

1. [RP-213599](http://www.3gpp.org/ftp//tsg_ran/TSG_RAN/TSGR_94e/Docs//RP-213599.zip), “New SI: Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface”, TSG RAN, RAN#94-e, Dec. 2021
2. [RP-221348](http://www.3gpp.org/ftp//tsg_ran/TSG_RAN/TSGR_96/Docs//RP-221348.zip), “Revised SID: Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface”, TSG RAN, RAN#96, Budapest, Hungary, June 2022
3. RP-230049, “Status of time budgets for RAN1/2/3/4 ongoing and new WIs/SIs after RAN #99”, RAN1 Chair (Samsung), RAN2 Chair (MediaTek), RAN3 Chair (ZTE), RAN4 Chair (Huawei), RAN#99, Rotterdam, Netherlands, March 2023
4. [R1-2205021](http://www.3gpp.org/ftp//tsg_ran/WG1_RL1/TSGR1_109-e/Docs//R1-2205021.zip), “Work plan for Rel-18 SI on AI and ML for NR air interface”, Qualcomm Inc., RAN1#109-e, May 2022
5. [R2-2210677](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_119bis-e/Docs//R2-2210677.zip), “RAN2 Work Plan for Rel-18 SI on AI/ML for NR air interface”, Ericsson, Qualcomm Inc., RAN2#119bis-e, October 2022
6. [RP-220534](http://www.3gpp.org/ftp//tsg_ran/TSG_RAN/TSGR_95e/Docs//RP-220534.zip), “Status report for SI Study on AI/ML for NR air interface; rapporteur: Qualcomm”, RAN1, RAN#95e, March 2023
7. [RP-221347](http://www.3gpp.org/ftp//tsg_ran/TSG_RAN/TSGR_96/Docs//RP-221347.zip), “Status report for SI Study on AI/ML for NR air-interface; rapporteur: Qualcomm” RAN#96, Budapest, Hungary, June 2022
8. [RP-222486](http://www.3gpp.org/ftp//tsg_ran/TSG_RAN/TSGR_97e/Docs//RP-222486.zip), “Status report for SI Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface; rapporteur: Qualcomm”, RAN1, RAN#97e, September 2022
9. [RP-223127](http://www.3gpp.org/ftp//tsg_ran/TSG_RAN/TSGR_98e/Docs//RP-223127.zip), “Status report for SI Study on AI/ML for NR air interface; rapporteur: Qualcomm”, RAN1, RAN#98e, December 2022
10. [RP-230762](http://www.3gpp.org/ftp//tsg_ran/TSG_RAN/TSGR_99/Docs//RP-230762.zip), “Status report for SI Study on AI/ML for NR air interface; rapporteur: Qualcomm”, RAN1, RAN#99, Rotterdam, Netherlands, March 2023
11. [TR 37.817](https://www.3gpp.org/ftp/Specs/archive/37_series/37.817/37817-h00.zip), “Study on enhancement for Data Collection for NR and EN-DC”, v17.0.0, April 2022
12. [RP-230295](http://www.3gpp.org/ftp//tsg_ran/TSG_RAN/TSGR_99/Docs//RP-230295.zip), “Status report for WI on Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN”, RAN3, RAN#99, Rotterdam, Netherlands, March 2023
13. [R1-2304168](http://www.3gpp.org/ftp//tsg_ran/WG1_RL1/TSGR1_112b-e/Docs//R1-2304168.zip), “Session notes for 9.2 (Study on AI/ ML for NR air interface)” Ad-hoc Chair (CMCC), RAN1#112bis-e, April 2023

# Annex A: Full list of RAN2 Agreements

Below the main agreements, observations and assumptions captured in the different RAN2 meeting discussions.

## RAN2#119bis-e (October 10 – 19, 2022)

Some initial Assumptions on the work:

- Assume that RAN2’s work can be somewhat split: A) use-case-centric configuration, signalling and control procedures, B) management of data and AI/ML models (where part of discussion may overlap between use cases).

- Assume that e.g. for the management of data and AI/ML models, RAN2 could start by focusing on data collection, model transfer, model update, model monitoring and model selection/(de)activation/switching/fallback (to the extent needed), whether UE capabilities has a role in this.

- Chair assumes that we will input on various aspects when the time is right, and e.g. postpone things that obviously need R1 decisions, but there could be some rare exception.

### AIML methods

* Assume that R2 will reuse terminology defined by R1 to the extent possible/reasonable
* Observation: the collaboration levels definitions doesn’t really clarify what is required, more work is needed
* R2 assumes that for the existing (under discussion) AI/ML use cases, proprietary models may be supported and/or open format may be supported (and maybe RAN2 doesn’t have to further elaborate on this assumption).
* R2 assumes that from Management or Control point of view mainly some meta info about a model may need to be known, details FFS.
* R2 assumes that a model is identified by a model ID. Its usage is FFS.
* General FFS: AIML Model delivery to the UE may have different options, Control-plane (multiple subvariants), User Plane, can be discussed case by case.

## RAN2#120 (Toulouse, France, November 14 – 18, 2022)

### AIML methods

* R2 assumes that model ID can be used to identify which AI/ML model is being used in LCM including model delivery.
* R2 assumes that model ID can be used to identify a model (or models) during model selection/activation/deactivation/switching (can later align with R1 if needed).
* For model transfer/delivery for AI/ML models (for the target use cases of this SI), RAN2 to study CP-based, UP-based solutions

### Use case specific aspects

* RAN2 scope includes procedures, protocols, and signaling for two-sided CSI use case(s), e.g.
1. Ensuring UE and gNB side models are configured / applied based on their applicable configurations / scenarios.
2. Ensuring that models are matched properly at both UE and gNB sides, i.e., when a CSI encoder is used at the UE corresponding CSI decoder is used at the gNB
3. Achieving simultaneous (de)activation and switching of the two-sided model

## RAN2#121 (Athens, Greece, February 27 – March 3, 2023)

### AIML methods

Data Collection

*Proposal 1 RAN2 to simultaneously focus on studying data collection solutions for both NW- and UE-sided AIML models, including assistance signalling and (dataset) reporting from the concerning entity.*

*Proposal 2 Study RAN2 implications of data collection for all concerning LCM purpose, e.g., model training/monitoring/selection/update/inference/etc.*

*Proposal 3 RAN2 to separately analyse the data collection requirements and solutions for the different LCM purposes. FFS if general frameworks/solutions could be adopted.*

*Proposal 4 Wait for RAN1 requirements before discussing specific data collection solutions for use cases and for the related (LCM) procedures. In the meantime, RAN2 can summarize the implementation of existing frameworks while focusing on different performance metrics.*

*Proposal 5 When summarizing the different data collection frameworks, RAN2 can start by considering the following metrics: a) the content of the data, b) the data size, c) latency and periodicity, d) signalling, entities involved, and configuration aspects. FFS on how to handle security/privacy.*

*Proposal 6 Consider the following existing frameworks as starting points to be considered for data collection: SON & MDT, UE assistance information, RRM measurement reports, CSI reporting framework, LPP Provide location information. FFS whether other frameworks should be discussed.*

*Proposal 7 Upon receiving specific (RAN1) requirements, RAN2 to decide whether the existing frameworks can be reused/extended, or whether a new framework is required.*

*Proposal 8 For data collection, RAN2 will simply keep progressing and will inform of concerning agreements to RAN1 when necessary.*

* P1-P8 are loosely endorsed with the understanding that we can also go beyond, e.g. analyse other methods.

Rapporteur’s Note: The following agreement is referring to [*R2-2300708*](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121/Docs//R2-2300708.zip):

* The table in this doc is endorsed as starting point

Rapporteur’s Note: The table in [*R2-2300708*](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121/Docs//R2-2300708.zip) (see agreement just above) led to a further iteration in [R2-2302286](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121/Docs//R2-2302286.zip) and the following set of agreements:

* Endorse the table as a starting point (e.g. can add more columns if needed later, modify, add rows etc). Content shall be interpreted as current content.
* Chair: There is significant support to aim for evaluating the data collection methods per LCM purpose

Model Transfer

* We Use the wording “model transfer/delivery”
* model delivery that serves the use cases in the SI is within RAN2 scope, regardless other aspects.
* Agreed:

Aim to at least analyze the feasibility and benefits of model/transfer solutions based on the following:

Solution 1a: gNB can transfer/deliver AI/ML model(s) to UE via RRC signalling.

Solution 2a: CN (except LMF) can transfer/deliver AI/ML model(s) to UE via NAS signalling.

Solution 3a: LMF can transfer/deliver AI/ML model(s) to UE via LPP signalling.

Solution 1b: gNB can transfer/deliver AI/ML model(s) to UE via UP data.

Solution 2b: CN (except LMF) can transfer/deliver AI/ML model(s) to UE via UP data.

Solution 3b: LMF can transfer/deliver AI/ML model(s) to UE via UP data.

Solution 4: Server (e.g. OAM, OTT) can transfer/delivery AI/ML model(s) to UE (e.g. transparent to 3GPP).

**Table: relations between solutions and use cases**

|  |  |
| --- | --- |
| **Solutions** | **Applicable use cases** |
| Solution 1a, 1b | CSI feedback enhancementBeam managementNote: No specific considerations for Positioning accuracy enhancement for Solution 1a and 1b. |
| Solution 2a, 2b | CSI feedback enhancementBeam managementNote: No specific considerations for Positioning accuracy enhancement for Solution 2a and 2b. |
| Solution 3a, 3b | Positioning accuracy enhancement |
| Solution 4 | CSI feedback enhancementBeam managementPositioning accuracy enhancement |

Note: the solutions use case relation is preliminary (work in progress), and the purpose is to have better understanding on what to further analyse

Chair think that in general, we may need to understand what issues are expected, e.g. Loosely Expect that time/latency from trigger to get a new model and until is downloaded and operational may be an issue, expect some other issue (in certain circumstances) and so on …

Rapporteur’s Note: The following agreement is referring to [R2-2302268](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121/Docs//R2-2302268.zip):

* The table can serve as starting point for continued discussion (but contains some parts that seems non consensus, e.g. delta configuration).

Model ID and UE cap

* RAN2 assumes that Model ID is unique “globally”, e.g. in order to manage test certification each retrained version need to be identified.

General

* R2 may consider including the existing EVEX framework for this SI, FFS exactly what this means, can discuss next meeting.

## RAN2#121bis-e (April 17 – 26, 2023)

### AIML methods

* R2 will deprioritize aspects of on-line/real-time training for the whole SI (unless R1 identifies that it is needed for one of the studied use cases).

#### Architecture General

* FFS if For UE capability for AIML methods we use the UE capability mechanisms as defined for RRC reported and LPP reported capabilities.
* For the CSI compression and beam management use cases, model/function selection/(de)activation/switching/fallback can be UE-initiated or gNB-initiated. FFS how the different cases are different (e.g. applicability to UE-sided vs network sided model).
* For the positioning use case, model/function selection/(de)activation/switching/fallback can be UE-initiated or LMF-/ gNB-initiated. FFS how the different cases are different (e.g. applicability to UE-sided vs network sided model).
* R2 assumes that Information such as FFS:vendor info, applicable conditions, model performance indicators, etc. may be required for model management and control, and should, as a starting point, be part of meta information.
* The general AI/ML framework consist of, (i) Data Collection, (ii) Model Training, (iii) Model Management, (iv) Model Inference, and (v) Model Storage.

Chair: the following was almost agreed (leave it FFS for now): AI/ML functional architecture in Figure 1 in [R2-2303674](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121bis-e/Docs//R2-2303674.zip) is the baseline with the modification that Performance Monitoring is changed to Model Mgmt / Performance Monitoring. It is noted that the exact interactions may need some modification depending on how each piece of functionality is specified**.**

* Model ID can be used to identify model or models for the following LCM purposes:

model selection/activation/deactivation/switching (or identification, if that will be supported as a separate step).

(e.g. for so called “model ID based LCM”)

* If model transfer/delivery is supported, model ID can be used for model transfer/delivery LCM purpose.
* How to achieve globality of the Model ID is FFS.

Initial discussion in RAN2: the following global unique model ID definition directions can be considered as a starting point:

Direction1: Pre-defined/hard-coded global unique model ID

Direction3: Assigned global unique model ID via specific ID management node.

Note: Other global unique model ID definition is not precluded.

Model ID structure, if any, is FFS

Chair: companies can also consider the remaining proposals and proposed open issues for later discussions.

Rapporteur’s Note: The chair’s observation above is referring to proposals and open issues in [R2-2304195](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121bis-e/Docs//R2-2304195.zip).

#### Data Collection

* Extend the previously endorsed table with 3 columns: Inference, Monitoring and Training, and explain in free text the applicability of the data collection method to the LCM purpose and the use case(s).
* Observation: RAN2 may need to consider enhancements for AIML to existing functionality for data collection, e.g. for timing control (e.g. for MDT/RRM).

Rapporteur’s Note: The following set of agreements relate to [R2-2304541](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121bis-e/Docs//R2-2304541.zip).

* P1: RAN2 to understand/determine/capture requirements of data collection for the LCM functionalities and document the results. FFS on the exact presentation format. Expect RAN1 to provide some related information.
* P2: RAN2 to capture the analysis (see P1 above) separately for the use-cases, i.e., CSI feedback enhancement, beam management and positioning enhancement. FFS how we do the formatting/presentation of the results.
* P3: Study the applicability (and limitations) of each identified data collection framework for each of the identified LCM purposes, i.e., inference, monitoring and (offline) training. FFS how we do the formatting/presentation of the results.
* P4: With more progress on architectural discussion, consider the suitability of each identified data collection framework for the termination points and mapping with the location of LCM purposes/functions (inference, monitoring, (offline) training)

- Model sidedness (UE side, NW side, two sided) FFS

- Use case mapping FFS

* P5: RAN2 to modify the previously endorsed table by adding 3 additional columns: inference; monitoring and (offline) training. Whether to, and how to further restructure the table is FFS.

Rapporteur’s Note: The following chair comments regarding EVEX where based from online discussion on [R2-2302954](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121bis-e/Docs//R2-2302954.zip).

Chair: There is some support to add EVEX as an option, but there is a lot of concerns. Majority of companies seems to have concerns.

Chair: Maybe the vivo proposal was too wide: Proposal: Add EVEX (or modified EVEX if needed) as one potential option for collection of data for training for UE side models.

- Huawei, ZTE, OPPO, CMCC, Ericsson and Apple object

# Annex B: RAN2 agreements organized by topic

For better readability, the Rapporteur has categorized below the agreements and observations by topic.

## General

* Assume that R2 will reuse terminology defined by R1 to the extent possible/reasonable
* R2 will deprioritize aspects of on-line/real-time training for the whole SI (unless R1 identifies that it is needed for one of the studied use cases).
* The general AI/ML framework consist of, (i) Data Collection, (ii) Model Training, (iii) Model Management, (iv) Model Inference, and (v) Model Storage.
* FFS if For UE capability for AIML methods we use the UE capability mechanisms as defined for RRC reported and LPP reported capabilities.

## Model identification and related meta data

* R2 assumes that for the existing (under discussion) AI/ML use cases, proprietary models may be supported and/or open format may be supported (and maybe RAN2 doesn’t have to further elaborate on this assumption).
* R2 assumes that from Management or Control point of view mainly some meta info about a model may need to be known, details FFS.
* R2 assumes that a model is identified by a model ID. Its usage is FFS.
* R2 assumes that model ID can be used to identify which AI/ML model is being used in LCM including model delivery.
* R2 assumes that model ID can be used to identify a model (or models) during model selection/activation/deactivation/switching (can later align with R1 if needed).
* RAN2 assumes that Model ID is unique “globally”, e.g. in order to manage test certification each retrained version need to be identified.
* R2 assumes that Information such as FFS:vendor info, applicable conditions, model performance indicators, etc. may be required for model management and control, and should, as a starting point, be part of meta information.
* Model ID can be used to identify model or models for the following LCM purposes:

model selection/activation/deactivation/switching (or identification, if that will be supported as a separate step).

(e.g. for so called “model ID based LCM”)

* If model transfer/delivery is supported, model ID can be used for model transfer/delivery LCM purpose.
* How to achieve globality of the Model ID is FFS.

Initial discussion in RAN2: the following global unique model ID definition directions can be considered as a starting point:

Direction1: Pre-defined/hard-coded global unique model ID

Direction3: Assigned global unique model ID via specific ID management node.

Note: Other global unique model ID definition is not precluded.

Model ID structure, if any, is FFS

## Data collection

*Proposal 1 RAN2 to simultaneously focus on studying data collection solutions for both NW- and UE-sided AIML models, including assistance signalling and (dataset) reporting from the concerning entity.*

*Proposal 2 Study RAN2 implications of data collection for all concerning LCM purpose, e.g., model training/monitoring/selection/update/inference/etc.*

*Proposal 3 RAN2 to separately analyse the data collection requirements and solutions for the different LCM purposes. FFS if general frameworks/solutions could be adopted.*

*Proposal 4 Wait for RAN1 requirements before discussing specific data collection solutions for use cases and for the related (LCM) procedures. In the meantime, RAN2 can summarize the implementation of existing frameworks while focusing on different performance metrics.*

*Proposal 5 When summarizing the different data collection frameworks, RAN2 can start by considering the following metrics: a) the content of the data, b) the data size, c) latency and periodicity, d) signalling, entities involved, and configuration aspects. FFS on how to handle security/privacy.*

*Proposal 6 Consider the following existing frameworks as starting points to be considered for data collection: SON & MDT, UE assistance information, RRM measurement reports, CSI reporting framework, LPP Provide location information. FFS whether other frameworks should be discussed.*

*Proposal 7 Upon receiving specific (RAN1) requirements, RAN2 to decide whether the existing frameworks can be reused/extended, or whether a new framework is required.*

*Proposal 8 For data collection, RAN2 will simply keep progressing and will inform of concerning agreements to RAN1 when necessary.*

* P1-P8 are loosely endorsed with the understanding that we can also go beyond, e.g. analyse other methods.

Rapporteur’s Note: The following agreement is referring to [*R2-2300708*](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121/Docs//R2-2300708.zip):

* The table in this doc is endorsed as starting point

Rapporteur’s Note: The table in [*R2-2300708*](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121/Docs//R2-2300708.zip) (see agreement just above) led to a further iteration in [R2-2302286](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121/Docs//R2-2302286.zip) and the following set of agreements:

* Endorse the table as a starting point (e.g. can add more columns if needed later, modify, add rows etc). Content shall be interpreted as current content.
* Chair: There is significant support to aim for evaluating the data collection methods per LCM purpose
* R2 may consider including the existing EVEX framework for this SI, FFS exactly what this means, can discuss next meeting.
* Extend the previously endorsed table with 3 columns: Inference, Monitoring and Training, and explain in free text the applicability of the data collection method to the LCM purpose and the use case(s).
* Observation: RAN2 may need to consider enhancements for AIML to existing functionality for data collection, e.g. for timing control (e.g. for MDT/RRM).
* P1: RAN2 to understand/determine/capture requirements of data collection for the LCM functionalities and document the results. FFS on the exact presentation format. Expect RAN1 to provide some related information.
* P2: RAN2 to capture the analysis (see P1 above) separately for the use-cases, i.e., CSI feedback enhancement, beam management and positioning enhancement. FFS how we do the formatting/presentation of the results.
* P3: Study the applicability (and limitations) of each identified data collection framework for each of the identified LCM purposes, i.e., inference, monitoring and (offline) training. FFS how we do the formatting/presentation of the results.
* P4: With more progress on architectural discussion, consider the suitability of each identified data collection framework for the termination points and mapping with the location of LCM purposes/functions (inference, monitoring, (offline) training)

- Model sidedness (UE side, NW side, two sided) FFS

- Use case mapping FFS

* P5: RAN2 to modify the previously endorsed table by adding 3 additional columns: inference; monitoring and (offline) training. Whether to, and how to further restructure the table is FFS.

## Model transfer/delivery

* General FFS: AIML Model delivery to the UE may have different options, Control-plane (multiple subvariants), User Plane, can be discussed case by case.
* For model transfer/delivery for AI/ML models (for the target use cases of this SI), RAN2 to study CP-based, UP-based solutions
* We Use the wording “model transfer/delivery”
* model delivery that serves the use cases in the SI is within RAN2 scope, regardless other aspects.
* Agreed:

Aim to at least analyze the feasibility and benefits of model/transfer solutions based on the following:

Solution 1a: gNB can transfer/deliver AI/ML model(s) to UE via RRC signalling.

Solution 2a: CN (except LMF) can transfer/deliver AI/ML model(s) to UE via NAS signalling.

Solution 3a: LMF can transfer/deliver AI/ML model(s) to UE via LPP signalling.

Solution 1b: gNB can transfer/deliver AI/ML model(s) to UE via UP data.

Solution 2b: CN (except LMF) can transfer/deliver AI/ML model(s) to UE via UP data.

Solution 3b: LMF can transfer/deliver AI/ML model(s) to UE via UP data.

Solution 4: Server (e.g. OAM, OTT) can transfer/delivery AI/ML model(s) to UE (e.g. transparent to 3GPP).

**Table: relations between solutions and use cases**

|  |  |
| --- | --- |
| **Solutions** | **Applicable use cases** |
| Solution 1a, 1b | CSI feedback enhancementBeam managementNote: No specific considerations for Positioning accuracy enhancement for Solution 1a and 1b. |
| Solution 2a, 2b | CSI feedback enhancementBeam managementNote: No specific considerations for Positioning accuracy enhancement for Solution 2a and 2b. |
| Solution 3a, 3b | Positioning accuracy enhancement |
| Solution 4 | CSI feedback enhancementBeam managementPositioning accuracy enhancement |

Note: the solutions use case relation is preliminary (work in progress), and the purpose is to have better understanding on what to further analyse

Rapporteur’s Note: The following agreement is referring to [R2-2302268](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_121/Docs//R2-2302268.zip):

* The table can serve as starting point for continued discussion (but contains some parts that seems non consensus, e.g. delta configuration).

## Use case specific agreements

### CSI and Beam Management

* RAN2 scope includes procedures, protocols, and signaling for two-sided CSI use case(s), e.g.
1. Ensuring UE and gNB side models are configured / applied based on their applicable configurations / scenarios.
2. Ensuring that models are matched properly at both UE and gNB sides, i.e., when a CSI encoder is used at the UE corresponding CSI decoder is used at the gNB
3. Achieving simultaneous (de)activation and switching of the two-sided model
* For the CSI compression and beam management use cases, model/function selection/(de)activation/switching/fallback can be UE-initiated or gNB-initiated. FFS how the different cases are different (e.g. applicability to UE-sided vs network sided model).

### Positioning

* For the positioning use case, model/function selection/(de)activation/switching/fallback can be UE-initiated or LMF-/ gNB-initiated. FFS how the different cases are different (e.g. applicability to UE-sided vs network sided model).