**3GPP TSG-RAN3 Meeting #114-e R3-215908**

**E-meeting, 1 - 11 Nov 2021**

**Agenda item:** 18.2

**Source:** Deutsche Telekom (moderator)

**Title:** CB: # AIRAN1\_Framework - Summary of email discussion

**Document Type:** Discussion and Approval

# 1 Introduction

**CB: # AIRAN1\_Framework**

**- Converge on the open issues on the AI functional framework**

**- Performance metrics (accuracy, validity time)?**

**- Update high level principles if needed**

**- Provide TPs if agreeable**

(DT - moderator)

Summary of offline disc [R3-215908](file:///D%3A%5CYang%20Xudong%5C3GPP%20meetings%5CRAN3-114%5CCB%5CCB%20%23%20AIRAN1_Framework%5CInbox%5CR3-215908.zip)

The offline discussion will comprise 2 phases:

* Phase 1: Try to get agreements on open and/or controversial issues on AI/ML functional framework and high-level principles.
	+ **Deadline: Nov 5th, 2021, Friday, 11 am UTC**
* Phase 2: Try to come up with a TP, if agreeable, based on the outcome of Phase 1.
	+ **Deadline: Nov 9th, 2021, Tuesday, 9 am UTC**

# 2 For the Chairman’s Notes

…

# 3 Discussion (Phase 1)

## 3.1 Outcome of RAN3#113-e meeting on “Functional framework for RAN intelligence”

At RAN3#113-e meeting, TR 37.817 [1] was updated covering amongst others the agreements achieved within e-mail discussion on high-level principles and definitions for AI/ML functional framework [2]. This version of TR 37.817 includes an updated figure of the functional framework under discussion (see Figure 1 below) as well as some explanations for the functional blocks and their interconnections (i.e., inputs/outputs) given in the figure.



**Fig. 1: Functional Framework for RAN Intelligence [10]**

Following open topics were raised during the e-mail discussion and in the online sessions (see [2] and the RAN3 chairman’s notes, respectively) or are stated as Editor’s notes in [1]:

* The definition of Model Deployment/Update (“Deploy or update an AI/ML model to Model Inference function.”) is still FFS.
* Whether to keep the Model Performance Feedback arrow from Model Inference to Model Training using a dashed line or together with some clarification text.
* FFS on whether model testing / generating of model performance metrics is performed in Model Inference.
* FFS if the study assumes single vendor environment, e.g., if model payload is proprietary and if the model deployment/update procedure is proprietary.
* FFS whether RAN is allowed to store user data and in which cases, the coordination across use cases need to be consistent.
* FFS whether and how to signal metrics (e.g., accuracy, uncertainty, etc.) and validity time together with or as part of the inference output.

## 3.2 Discussion on the open issues on high-level principles and AI/ML functional framework

## 3.2.1 Model Deployment/Update

The discussion on that topic is continuing from last RAN3 meeting. Many companies provided again an input on it for RAN3#114e. Based on those ([4], [5], [9], [10], [12] - [16]) there seems to be a general consent to keep Model Deployment/Update in the figure and to remove the FFS, as the AI/ML framework is illustrated from functional point of view, and does not imply any specific interface and specification work. Without the Model Deployment/Update interconnection the logic of AI/ML functionality covering also the relevant part of model lifecycle management (LCM) process is broken and does not provide the full picture. It is also clear that the specification of the model deployment/update process is out of scope of RAN3.

Some companies proposed to add a note to Model Deployment/Update description that details are not considered in the current Rel-17 SI [12] or that details are vendor proprietary [15].

[4] clarified that the initial model deployment is a mandatory process to be considered in the model LCM, but that the need for model update depends on use cases under consideration, i.e., it is just optional. To make that clear in the figure of the functional framework it is proposed to split current interconnection into 2 parallel parts (arrows) using a solid arrow for Model Deployment and a dashed one for Model Update as shown in following figure. An explanation for that description is additionally given in a TP in [4].



[8] proposed to introduce a new function called Model Management which is drawn between Model Training and Model Inference in the figure. Instead of Model Training this Model Management is responsible for performing Model Deployment/Update to Model Inference based on information received from Model Training (see below). In addition, a description for the Model Management is given to be added to Sec. 4.2 of TR 37.817.



**Question 1: Companies are kindly asked to provide feedback to Model Deployment/Update:**

1. **Do you agree to keep the Model Deployment/Update arrow in the figure on functional framework and to remove the FFS?**
2. **Do you see the need to add a NOTE as proposed in [12] and [15]? Which one do you prefer?**
3. **Do you agree to make a differentiation between Model Deployment and Model Update as described in [4]? If yes, any comments to the updated figure and proposed description to be added to Sec. 4.2 of TR 37.817?**
4. **Do you agree to add the new Model Management function as proposed in [8]? If yes, any comments to the updated figure and proposed description to be added to Sec. 4.2 of TR 37.817?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No to (1), (2), (3), and (4)** | **Comments** |
| Qualcomm | (1) Yes(2) No(3) No(4) Yes | There is no direct interworking between Model Training and Model Inference except in reinforcement learning. The interworking between them is coordinated by Model Management as shown in figure below. One intention of our framework figure is to analyse the signalling need. Without the model management, we cannot analyse the signalling impact correctly. Depending on use case and deployment scenario, the model management role may be taken by OAM, CU or other NE and then results in different signalling impact. 图示  描述已自动生成 |
| Huawei | Yes to (1)Yes to (2)No to (3)No to (4) | For (1), yes, we could keep this arrow in order to reflect a complete view of different function for AI/ML operation;For (2), as proponent, we of course prefer note in [15], we think offline are online should be differentiated;For (3), maybe not needed though technically nothing wrong, but they are two actions of the same attributes, and both of which are optional (for online training, this could even be done entirely inside RAN)For (4), this indeed makes the framework complicated and even ambiguous, model deployment/update could be actually part of management function. And this framework is mainly for function description while for management, this could be discussed in SA5 which may have another framework. |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## 3.2.2 Model Performance Feedback

The open issues on Model Performance Feedback are related to the aspect if such feedback is needed from Model Inference function to Model Training and if yes, if it is a mandatory or optional feedback, e.g., based on the use case implemented. This is also related to the question if model testing / generating of model performance metrics is performed in Model Inference function.

Many companies provided again an input on this open topic for RAN3#114e. Most companies proposed to keep the Model Performance Feedback in the figure of functional framework and to remove FFS ([4], [5], [8], [10], [12] - [14], [16]) as it is used to feedback, e.g. the effectiveness of a model and/or to trigger the AI/ML model retrain/update at Model Training in case the model performance degrades. As this may be dependent on the LCM approach selected or the use case under consideration, it is proposed to change the solid line of the arrow to a dashed line to make clear that it is only optional. In [8] again the Model Management as new function is introduced (see the corresponding figure in Sec. 3.2.1 of this SoD).

Three companies proposed to remove the arrow ([6], [7], [15]) as they stated that performance level information cannot be provided by Model Inference function without help of the Actor. In [6] it is mentioned that it is the Data Collection function to evaluate the performance of AI/ML models. Therefore, a change of the figure is proposed by drawing the arrow of Model Performance Feedback from Data Collection toward Model Training. In addition, a dashed arrow named also Output should be drawn from Model Inference toward Data Collection for prediction-based AI/ML (see below).



**Question 2: Companies are kindly asked to provide feedback to Model Performance Feedback:**

1. **Do you agree to keep the Model Performance Feedback arrow in the figure on functional framework, to use a dashed line for the arrow to depict the optionality, and to remove the FFS?**
	1. **If “yes” to (1), is there any preference with respect to the description to be added in Sec. 4.2 of TR 37.817 ([4], [10], [12] - [14], as well as [8] with inclusion of the Model Management function)?**
2. **If “no” to (1), do you see the need to adapt the figure as proposed in [6]?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No to (1) and (2)** | **Comments** |
| Qualcomm |  | The performance feedback should be sent to model management for model performance monitoring. Then, model management can decide to update the model, fallback to legacy algorithm or re-train the model. |
| Huawei | No to (1) and (2) | As we already discussed, to remove model performance feedback doesn’t mean we don’t need feedback, technically we think the actual performance is only available after we see outcome from actor. Then we also see no need to update the framework in [6]. |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## 3.2.3 Single vendor environment

There is still the discussion if the study assumes a single vendor environment, e.g., if model payload is proprietary and if the model deployment/update procedure is proprietary. It was agreed during last RAN3 meetings that the AI/ML model itself is vendor proprietary.

In [4] as input to RAN3#114-e it was clarified that this does not mean that the Model Deployment/Update process as part of model LCM has to be proprietary, too (which was proposed in [7]) and that there are no good reasons to restrict the model deployment process to single vendor approaches within the ongoing SI, as finally the responsibility for defining this process should be in OAM, i.e., it is related to SA5 work. Therefore, there is also no need for RAN3 to go into details.

In [12] it was observed that Model Deployment/Update of single vendor/proprietary models can be based on OAM without standardization or can be supported by standardized procedures. For multi-vendor environment, standardized procedures for Model Deployment/Update would need to be designed in a way to avoid ML Model exposure over the network interfaces, while at the same time enabling the recipient of an ML Model to understand how to execute it, but this would go beyond the scope of the Rel-17 SI.

**Question 3: Companies are kindly asked to provide feedback to the single vendor environment:**

1. **Do you see the need to explicitly state in TR 37.817 that the study assumes a single vendor environment for the functional framework in Rel-17 which also addresses functions/processes being out of scope of RAN3’s work?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No to (1)**  | **Comments** |
| Qualcomm | No | Single vendor environment is up to implementation and does not need standard discussion. If cannot complete in R17 SI, we can defer this to R18 SI. The deferring decision can be made in later phase on need. |
| Huawei | Yes and No | On one hand, we think the details of model and algorithm coupled with functions and process are out of scope of RAN3’s work; on the other hand, maybe there is also no need to explicitly mention single vendor, anyway standardization should not limit to single vendor, and the rest is up to vendors’ implementation and coordination.  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## 3.2.4 Data storage in the RAN

It is still FFS whether RAN is allowed to store user data and in which cases; the coordination across use cases need to be consistent.

In [15] it was observed that it is not RAN’s duty to provide database services, which are built on top of data storage function (basic part of a data server where a software program/platform is used to provide database services like storing, processing, and securing data, and data server contains the installation of the database system, from which database services are provided and consumed by other software programs or components). The Data Collection function doesn’t require a data storage function but is related to storage of temporary data only. The introduction of AI/ML operation at RAN side should not force RAN to provide a data base service function. As location of offline training which requires a large amount of data storage it was common understanding that OAM should be the suitable place. For online training it is quite similar as the existing function of measurement report, i.e., related data processing it is already supported by existing framework.

[4] stated that storage of large data volumes as e.g. required for (initial) offline training of AI/ML models should be avoided as much as possible in the RAN, but the feasibility of a RAN data repository function for AI/ML-based data analytics (see also [11]) should not be excluded from the study. The demand for it has to be evaluated based on use cases under consideration.

[9] noted that the network should be able to select most appropriate UEs to use AI/ML model for prediction, as it can help to reduce system computation complexity and reduce latency during performing model inference. The selection may be based on UEs’ QoS requirements, based on RAN measurements and/or based on indication from CN or UEs.

**Question 4: Companies are kindly asked to provide feedback to data storage in the RAN:**

1. **Do you see the need to explicitly exclude a data storage function (as defined in [15]) in the RAN from current scope of Rel-17 SI as there may only the need for temporary storage of data?**
2. **Do you agree to have a statement in the high-level principles in Sec. 4.2 that initial offline model training should be located out of the RAN domain, e.g., in OAM? Alternatively, this could be defined case-by-case dependent on the use case under consideration.**
3. **Do you see the need to coordinate with SA3 on security aspects for user data storage in the RAN?**
4. **Do you see the need to add a statement that the network may select the most appropriate UEs for AI/ML purposes based on certain criteria as mentioned in [9]?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No to (1), (2), (3), and (4)** | **Comments** |
| Qualcomm | (1), No(2) Yes(3) Not now(4) Yes |  |
| Huawei | Yes to (1) and (2), in general yes to (3), No to (4) | For (1), as we pointed out in our discussion paper, data storage is totally different temporary storing data, temporary storing data is already allowed today.For (2), we think it should be good to have such clarifications, at least offline training is out of RAN domain and our of RAN3 scope.For (3), in general, we think it should be good for SA3 to study/investigate whether there is any additional security risk when AI/ML operation is introduced in RAN. But as already said, we think data storage/data base service is not RAN function, but the security part of data storage/data base service (data server), if located in OAM, should be SA4 work to check. Technically, we even see that data server may even be located outside of 3GPP.For (4), considering conspiracy, we don’t think this is a good idea, even, we have not discussed if there is a need to have additional task over a specific UE concerning AI/ML operation. Let’s focus on the solution to agree use cases for the moment.  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## 3.2.5 Metrics and validity time

It is still FFS whether and how to signal metrics (e.g., accuracy, uncertainty, etc.) and validity time together with or as part of the inference output. Several companies provided input on that topic to RAN3#114-e.

In [3] it was stated that after AI/ML model is initially trained/validated or after it’s updated/retrained, the Model Training function should generate model performance metrics calculated using the identified test dataset(s). The calculated model performance metrics (differentiated according to accuracy-, reliability/robustness-, and uncertainty related metrics) should be delivered to the functions that subscribe to receive such information. After AI/ML model deployment, the Model Inference function should generate the model performance metrics which are calculated using newly collected field data. The calculated model performance metrics should be delivered to the functions that subscribe to receive such information, e.g., the Actor.

[7] is proposing to include a validity time together with a Model Inference prediction output. Furthermore, it is proposed not to signal accuracy metrics together with or as part of the Model Inference output. Standardization of such metrics should be avoided as they are strongly problem dependent and there may be the risk that they are only applicable for a limited set of implementations. Also, it is questionable whether the Model Inference function is able to provide an estimation of accuracy for all the generated predictions.

[9] noted that an accuracy level for a tested AI/ML model should be provided from Model Training to Model Inference when deploying/updating that model. The Model Inference function should propagate the accuracy level to the Actor function. It may also generate and calculate the confidence level between predicted output and actual value of the system. Model testing/generating of model performance metrics is not supported in Model Inference, therefore the direct performance feedback/metrics (e.g., confidence level) from Model inference to Model training is not needed. But on the other hand, Model Inference Assistance Information from Model Inference to Model Training function may be considered, indicating capability (HW, storage, etc.) of the network entity holding the Model Inference.

In [12] it is proposed that the level of accuracy of the inference and validity time should not be discussed in the AI/ML framework to keep it simple, general, and independent from AI/ML algorithms and use cases but should be discussed based on case-by-case basis.

[13] stated that whether validity time and accuracy are needed needs to be discussed case-by-case. There is no need to explicitly indicate them in the AI functional framework or high-level general principles. There is no need to carry the accuracy indication information in the output from Model Inference as accuracy level indication cannot be calculated/measured until the action is executed based on the prediction provided by Model Inference function.

[14] proposed to set validity time (i.e., “best time period or time point” for the inference result) as well as accuracy as additional information provided by the Model Inference function together with the inference output.

[15] is stating that technically, accuracy is to describe the precision of a model training, while confidence is to indicate the probability of inference output approaching the reality; but they seem to be used for the model training or inference to evaluate its performance by itself. There is no need to include the validity time together with the output information every time. Instead, it can be considered on a case-by-case manner when the use case and solution are discussed.

**Question 5: Companies are kindly asked to provide feedback to metrics and validity time:**

1. **Do you agree that metrics (which may be related to accuracy, reliability/robustness, or uncertainty) and validity time do not have to be explicitly listed and described in the high-level principles and/or the functional framework but should be covered in the use case descriptions?**

**Note 1: In Sec. 4.2 of TR 37.817 it is already stated that “whether there is any standardization impact and what is the standardization impact are discussed in clause 5”.**

**Note 2: In case of “yes” to (1) a simple note at the description of Output of Model Inference function should be sufficient referring to the use case descriptions in TR 37.817.**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No to (1)** | **Comments** |
| Qualcomm | Yes |  |
| Huawei | Yes, but… | There is no need to explicitly to list and describe in the high-level principles and/or the functional framework, and even it is not clear the benefits of introducing such parameters for a specific use case… |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## 3.2.6 Other issues

[11] proposes to align the architectural aspects of the AI/ML functional framework to that derived by SA2 for NWDAF introducing new functionalities like RDAF (RAN Data Analytics Function), RDRF (RAN Data Repository Function), and RDCF (RAN Data Coordination Function). To avoid introducing new interfaces in this first release of RAN AI/ML, the new entities can be thought as functions of a gNB. Therefore, the proposal is to describe those functionalities in an Annex of TR 37.817 by mapping them to the functional framework:

Annex A: Architecture of RAN AI/ML

The AI/ML functions given for the functional framework for RAN intelligence may be supported by network entities as shown in Figure A-1.



**Figure A-1: Functional mapping**

The Model Inference and online Model Training function can be supported by RDAF (RAN Data Analytics Function). The data collection function is carried out by RDRF (RAN Data Repository Function) for storage. Data collection to RDRF and distribution to RDAF (i.e., to Model Training/Inference functions) are coordinated by RDCF (RAN Data Coordination Function). The Actor role is taken by gNB-CU, gNB-DU.

The new network entities are thought as part of gNB in this release of the specification.

[12] proposes to introduce a cost per AI/ML measurement request sent from an entity producing measurements to Model Training. The cost amounts to a number of units from a total budget. A network entity (e.g., a gNB, gNB-CU, gNB-DU) providing AI/ML measurements can indicate to Model Training (e.g., OAM, gNB, gNB-CU) a budget for providing those as well as a cost for the measurements it can provide. AI/ML measurement producer provides measurements to Model Training under the condition that its budget for AI/ML measurements is not exceeded. A TP is provided to add a bullet point on that to high level principles in Sec. 4.1 of TR 37.817:

“An entity providing measurements for AI/ML may indicate to the Model Training its available budget for providing those measurements, as well as a cost for each measurement it can provide. The entity responsible for Model Training can use this information to determine which measurements it should request so that the budget at the measurement producer is not exceeded.”

**Question 6: Companies are kindly asked to provide feedback to the 2 other issues mentioned before:**

1. **Do you agree to add an Annex on alignment of architectural aspects of the AI/ML functional framework to that derived by SA2?
If yes, do you see any modifications needed to the proposed TP given in [11]?**
2. **Do you agree to add the bullet point proposed in [12] on cost budget for Model Training** **to high level principles in Sec. 4.1 of TR 37.817?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No to (1) and (2)** | **Comments** |
| Qualcomm | 1. Yes
 | Capturing the TP in [11] into Annex is useful. It provides a good reference for how the framework is supported/implemented.  |
| Huawei | No to (1) and (2) | We are not sure why we need to align RAN’s description with SA2, in SA2, network entity is network function based while this is not RAN’s approach; while for cost, in our understanding, AI/ML is just a way of achieving RRM purpose for which we use kind of if-else logic today, it is strange that RAN node would use a cost indication to show its capacity of dealing with mobility handling? |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# 4 Discussion (Phase 2)

**…**

# 5 References

1. TR 37.817: Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Study on enhancement for Data Collection for NR and EN-DC; Rel-17, V0.3.0 (2021-08).
2. R3-214220: Summary of offline discussion on AI RAN general framework; Lenovo, Motorola Mobility (moderator); RAN3#113-e.
3. R3-214733: RAN Intelligence Framework – Further Discussion on Model Testing and Model Performance Evaluation (Futurewei)
4. R3-214796: Discussion on open issues within AI/ML-based functional framework for RAN intelligence (Deutsche Telekom)
5. R3-214951: Proposed resolutions to open issues on functional framework (NEC)
6. R3-215055: Discussion on the model performance feedback arrow (CATT)
7. R3-215237: (TP for SON BL CR for TR 37.817) Framework for RAN intelligence (Ericsson)
8. R3-215244: Model management in AI/ML framework (Qualcomm Incorporated)
9. R3-215268: High level principle and Functional Framework of AI/ML enabled NG-RAN Network (Intel Corporation)
10. R3-215330: Remaining issues on AI functional framework (Lenovo, Motorola Mobility)
11. R3-215464: Architecture for RAN AI/ML (Deutsche Telekom, Qualcomm Incorporated, T-Mobile USA)
12. R3-215477: (TP for TR 37.817) Open points on AI/ML Framework and Data Collection Discussions (Nokia, Nokia Shanghai Bell)
13. R3-215523: Further discussion on AI Functional Framework for RAN intelligence (ZTE Corporation, China Unicom)
14. R3-215561: Discussion on Functional Framework and High-Level Principles (Samsung, Verizon Wireless)
15. R3-215664: Further discussions on remaining open issues about general principles and frame work (Huawei)
16. R3-215694: Remaining aspect of AI framework (CMCC)

# 6 Conclusion, Recommendations [if needed]

If needed.