**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:\Yang%20Xudong\3GPP%20meetings\RAN3-114\CB\CB%20%23%20AIRAN1_Framework\Inbox\R3-215908.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. |
| Deutsche Telekom | (1) Yes  (2) Yes but …  (3) Yes  (4) No | (1) Otherwise the AI/ML model LCM process showed in a simplified way in the figure would be broken.  (2) Ok to add note but we don’t see the content as proposed in [12] and [15] but prefer statements as proposed in [4].  (3) We see it useful to make the differentiation in the figure and to cover the differences also in the description in Sec. 4.2 of TR 37.817.  (4) In principle, Qualcomm is right, but as model management is out of RAN3’s scope (🡪 SA5) we don’t see the need to extend the figure by the related function. As already mentioned, it should be sufficient to clarify in Sec. 4.2 e.g. via a note that Model Deployment/Update are not related to RAN3’s scope (see also the LS sent to SA5). |
| Lenovo, Motorola Mobility | Yes: (1) (3)  No: (2) (4) | (1) without model deployment/update, the functional framework is simply broken  (2) we don’t think it is necessary to put such restriction from spec point of view. We can clarify the payload is vendor proprietary as suggested in [4] if helpful.  (3) it sounds a good idea, we can work on the exact wording in phase 2.  (4) in our understanding model management is part of model training function, which is also very implementation dependent. |
| Samsung | Yes: (1)  No: (2) (4)  No strong view: (3) | (1) To make the framework logical and workable, model deployment/update arrow need to be kept.  (2) The study is still in the initial stage. So we prefer to not add such limitation at this stage.  (3) No strong view for it. Combined one or separate ones are OK.  (4) The framework is just provide the guideline for use case study. It is better to keep it brief. And it seems the function of model management is up to implementation, which is out of scope. |
| T-Mobile USA | 1. Yes 2. No strong view 3. Yes 4. No | 1. Basic logic framework to make it work 2. But prefer not to add too many restrictions at this stage 3. Yes, however we could see this be entirely implementation specific and covered by the existing function box 4. Should keep the scope withing RAN3 |
| Nokia | (1), (2): Yes  (3), (4): No | (1),(2): We acknowledge the importance of Model Deployment Update in completing the model lifecycle so we can remove the FFS, but we support to clarify in the TR that details of Model Deployment Update are not in the scope of Rel. 17 as in [12].  (3): We don’t need to split the arrow in two to indicate a mandatory Model Deployment and an Optional Model Update. It is simpler to show a single arrow.  (4): Model Management belongs to Training. There should not be a separate function introduced. |
| Verizon | (1) Yes  (2) No  (3) No strong view  (4) No | (1) Yes, gives a full picture, allows LCM of AI/ML apps  (2) No need to put limitations as it could preclude LCM as well as interoperability  (3) No strong view on keeping together vs sepatae as long as we have these represented in the architecture  (4) It complicates the architecture. Also model management function is within scope of SA5. |
| Futurewei | 1. Yes 2. Yes but… 3. No 4. No | (2) We think the 2 proposed notes are too restrictive at this SI stage. If we have to add a note for “Model Deployment/Model Update”, then we suggest using “details of model deployment/update is implementation dependent”.  (3) Our view is that the current wording of “Model Deployment/Model Update” covers either “Model Deployment” or “Model Update” scenario and it does not necessary mean that “Model Update” always applies. It’s better to keep the diagram simple while a note could be added to indicate “Model Update” may depend on use case and AI/ML algorithm design.  (4) Our suggestion is to keep the diagram at high level and leave “Model Management” to vendor implementation. |
| NEC | 1. Yes 2. Maybe 3. Maybe 4. Maybe | 1. We support keeping Model Deployment/Update arrow in the functional framework Figure. 2. Interfaces between Model Training and Model Inference are out of scope of this study. This is described in general principles. If a similar note further clarifies this, we are OK with such note. 3. Proposal [4] could be a good compromise to resolve FFS. 4. This may be beneficial. |
| China Telecom | (1) Yes  (2) No  (3) Yes  (4) No | Model deployment/upgrade are very important steps to make the entire loop more reasonable, and FFS should be removed. And it is reasonable to distinguish whether the two operations are mandatory or optional by dashed or solid lines. |
| Intel | 1) Yes  2) No  3) Not necessary  4) No | For 2), the arrow of model deployment and update is captured to provide reader a basic understanding of LCM of AI/ML. Although defining the exact procedure of LCM is the scope of SA5, we see no harm to capture in RAN3 as a full picture.  For 3), in general, we agree with the intention of proposal in [4], we also think certain rules need to be considered for model update, e.g. per request from model inference or based on certain conditions which can trigger model update, for either offline training or online training. This also means model update could be optional based on the system requirement. However, from the functional framework point of view, we prefer to keep the arrow as solid for both deployment and update to keep a full picture. However, we can clarify model update is optional or per request, etc in the terminology description.  For 4), model management is different from the concept of other functional nodes. For example, data collection, model training, model inference, and actor are described from data/action/model point of view, which are belonging to either control plane or user plane. However, model management is management plane concept. Hence, we think it’s not suitable to be captured in functional framework. |
| CATT | (1) Yes  (2) Not needed.  (3) Not needed.  (4) No for now. | (2) and (3) are almost common understanding and thus no need to capture, especially for (3) which will further complicate this figure.  We have some sympathy on the reason for (4) as shown by Qualcomm, but the revised figure for TR 37.817 isn’t aligned with Qualcomm’s quotation, where more boxes (especially the “AI/ML Continuous Operation” box) are involved.  And if the “AI/ML Continuous Operation” box is combined into the “AI/ML Inference” box, why can’t the “AI/ML Model Management” box combined into the “AI/ML Training” box as well? |

## 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]. |
| Deutsche Telekom | (1) Yes | (1) We have to differentiate between RAN KPI feedback via the actor and model related information from inference to training which can be optional. This has been raised in many inputs to RAN3#114-e.  A detailed explanation can be created in 2nd round based on a merge of inputs from tdocs listed above, if it is finally agreed to keep the Model Performance Feedback in the figure.  As explained in our answer to Question 1 we don’t see the need to include the Model Management function. |
| Lenovo, Motorola Mobility | Yes: (1) | As explained by the moderator, we also regard model performance feedback optional and can be captured using a dash line. We can work on the exact wording in phase 2. |
| Samsung | Yes: (1) | It is better to keep model performance feedback from “Model inference” to “Model training” to transfer the model evaluation. A dash line or adding description can be used to show the optionality.  The detailed wording can be discussed in phase 2. |
| Nokia | (1): Yes | We support keeping Model Performance Feedback from Model Inference to Model Training, to show a complete workflow, but with a note that it is not in the scope of Rel. 17 SI. How Model Inference is capable to evaluate Model Performance is up to implementation. |
| Verizon | (1) Yes  (2) No/Yes | (1) This is an optional feedback that may be relevant in certain models and not in others. So prefer to use a dotted line.  (2) Ref. 6 has two proposals. “No” to model performance feedback from data collection to model training box. “Yes” to dash line of “output” from “Model inference” to “Data Collection” box as this is needed for model chaining. |
| Futurewei | Yes | We agree keeping the Model Performance Feedback arrow. We also agree that feedback from the Actor is required to calculate/derive the Model Performance (metrics). Calculating the proper metrics may be implementation-dependent; thus, we suggest letting the Model Inference function take such responsibility. There are 2 options:   1. Add a note to indicate that “feedback” information from the Actor is required at the “Model Inference” to calculate the Model Performance feedback. 2. Draw a dash-line from Data Collection to Model Inference to indicate that the Model Inference function will receive the feedback from the Actor (through Data Collection function).   We think option a) should be sufficient and this will keep the diagram at high level, but option b) is ok if we prefer making it clear on the diagram.  **Note:** the question does **NOT** completely address the **FFS for “model testing / generating of model performance metrics is performed in Model Inference function**”, specifically, the FFS item of whether “model testing” is performed in Model Inference function. As agreed in RAN3 #113e, section 4.1 of TR37.817 (High-level Principles) already specify that “An AI/ML model used in a Model Inference function has to be initially trained, validated **and tested** before deployment”. Thus, it is logical that the **“Model Training” function should perform model testing procedure** and generate model performance metrics as part of the model testing outcome (using testing dataset). After the model deployment (to the “Model Inference” function), model performance metrics can be monitored and generated based on newly collected real-world (feedback) data at inference stage as well by the Model Inference function (please see our feedback for question in 3.2.5). |
| NEC | Yes | We support keeping Model Performance Feedback in the functional framework. Proposal [4] could be a good compromise to resolve FFS. If Model Management is added, figure will need to be updated. |
| China Telecom | Yes to (1) | Model performance feedback is very necessary, and the model training function should notify the model inference and data collection modules to provide corresponding information according to their own training needs. According to the evaluation of model prediction accuracy or confidence in model training, model performance feedback is not mandatory. You can choose to optimize the model through feedback when the model prediction performance declines. |
| Intel | 1) Ok | If the arrow is kept, we prefer a general description:  Model Performance Feedback: it includes information to help Model Training function continue training AI/ML model which is optimized and suitable for deployment at Model Inference function. |
| CATT | (1) Acceptable but not preferred  (2) Yes | Anyhow we have to find a way for the following case:  The AI/ML model is a type of UE-associated prediction.  Model training locates at the gNB 1, and so does the inference.  The inference function generates an output for the UE at 10:00, predicting something w.r.t. of the UE at 10:10.  But at 10:01 the UE is handed over to gNB 2.  And of course gNB 1 need some information to be provided from gNB 2 after 10:10 in order to know whether its prediction comes true.  Should such information provision tagged with “Model Performance Feedback”? No it shouldn’t. It is not gNB 2 who evaluates the model performance, nor does it host the “Inference” function.  Should such information provision tagged with the ordinary “Feedback”? No it shouldn’t either. gNB 2 is nor the “Actor”.  So in the end we find such information provision not addressed in the chart. This is a bug, we have to fix it.  And this is why we think we should make the model performance evaluation function a part of the data collection function, i.e. in order to make the abovementioned provision a kind of data collection.  Nevertheless other solutions are also acceptable for us. |

## 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. |
| Deutsche Telekom | No | RAN3 has already clarified that AI/ML model and algorithms can be vendor proprietary, but processes for model LCM like model deployment or update are out of scope of RAN3 and therefore, we should not set any limitation in our study with respect to this (or following) release. |
| Lenovo, Motorola Mobility | No | In the current TR, it states clearly:   * “For the functions and data/information flows shown in the Figure 4.2-1, whether there is any standardization impact and what is the standardization impact are discussed in clause 5.”   Which is fair enough, we don’t need to discuss the single vendor environment for the functional framework. |
| Samsung | No | This issue is implementation related. So we prefer to not add the limitation. |
| Nokia | No | Algorithms are outside the scope of this study and are vendor proprietary. Also, our thinking is that the interfaces between Model Training and Model Inference should not be part of the current scope and we should prioritize Rel-17 study of interoperable mechanisms for data collection. |
| Verizon | No | Standards are there to support multi-vendor environment. So we should not place unnecessary limitation in the architecture from the beginning. If this cannot be done in R17, we can defer to R18. |
| Futurewei | No | We think the “single vendor” assumption is too restrictive. We agree that for multi-vendor environment, the “Model Deployment/Model Update” procedure has to be designed in a way to allow the recipient to use the deployed model to perform the inference, however, those belong to implementation details. |
| NEC | No | As mentioned in several papers, if interfaces between Model Training and Model Inference are not specified in this SI, this does not always mean that they are from the same vendor and other way around. |
| China Telecom | No | No need for explicit statement. Some details in specific use cases may involve the implementation of a single vendor. However, it seems that the general functional framework does not just apply to single vendor, and various issues such as vendor interoperability under this framework need further discussion. |
| Intel | No | As mentioned by companies’ contribution, inter-vendor environment is not precluded in SA2.  We agree that AI/ML model itself is vendor proprietary. However, supporting inter-vendor environment does not mean that AI/ML model or algorithm should be specified. Supporting inter-vendor environment means the signaling procedure of model/data transfer between different functional nodes should be specified. Considering we will also specify what data is needed for certain use case, AI/ML model itself does not need to be exposed. Only the required input/output data need to be transferred together with the model itself.  Hence, without interpreting AI/ML model itself, “Model Training” or “Model Inference” node can still work with the knowledge of input/output of the corresponding model. |
| CATT | No | It is anyhow possible that multiple vendors share the same AI/ML mechanism. |

## 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. |
| Deutsche Telekom | (1) No  (2) Yes, but …  (3) Not now  (4) No | (1) More discussion is needed about what is explicitly meant by a data storage function as proposed in [15].  (2) In principle, we are fine with such a statement, but it has to be considered that this would exclude some proposals made in the use case descriptions. Based on the out use case evaluation outcome we could consider that in the conclusions, if required an agreed.  (3) Can be done in later step if we have more clarification of use case impacts.  (4) This is part of the data preparation for Model Training or Inference functions and can be left to implementation. |
| Lenovo, Motorola Mobility | Maybe: (3)  No: (1)(2)(4) | (1) even now RAN is able to store some data which is useful for NW operation, right?  (2) this can be discussed under each use case, don’t need to describe online /offline training in the framework, since they are AI algorithm dependent.  (3) maybe yes, but we need to make it clear what kind of user data we are talking about and why the handling is different from legacy  (4) it seems an implementation problem, no need to capture in the spec. |
| Samsung | Maybe: (3)  No: (1)(2)(4) | (1) Maybe more discussion is required to clarify the data storage.  (2) The model training need to consider the model functionality, input/output availability and volume. Prefer to discuss it case by case.  (3) When the specific security issue is required to be solved during use case study, we can coordinate with SA3 then,  (4) It seems it is up to implementation. |
| Nokia | (1) yes  (2) not needed  (3) yes  (4) No | (1) The current Rel.17 study assumes that no changes to the current architecture are allowed. Introducing a RAN data storage entity would require changes to the current architecture and need for new interfaces. So, we do not support the existence of such storage function.  (2) We have captured already in the TR 37.817 that Training can be offline (in OAM) or online in the RAN for the different use cases of interest. So, in our view this is not needed to be repeated in the high-level principles.  (3) yes, when there is need for coordination  (4) It could be useful for the network to be able to identify UEs for ML using more refined selection criteria than cell-based criteria. However, in our view AI/ML in the UE should not be in the scope of this study. |
| Verizon | (1) No  (2) No  (3) May be  (4) Yes | (1) Question is vague and unclear. It is reasonable to assume that data storage functionality would be needed to support AI/ML in NG-RAN  (2) No need for such unnecessary restrictions  (3) When there is a security issue identified, RAN3 should coordinate with SA3  (4) UE specific AI/ML optimization can allow operators to tune RAN specific to a UE service needs and envoronment |
| Futurewei | 1. No, but… 2. No 3. Yes, maybe later 4. No | 1. In general, we agree what is proposed in [15] “AI/ML operation at RAN side should not force RAN to provide data base service function”. However, data storage function doesn’t imply it will be used to perform data base service. We should still give flexibility to allow vendors to implement their own solutions which may realize a data storage function for short term needs to perform AI/ML operations. 2. We think this can be handled on a per use case basis while we could add a note to indicate that offline training needs to consider data storage needs. 3. We think this should be handled on a per use case basis; thus, adding such blanket statement in the principles section may not be appropriate. |
| NEC | 1. No 2. No 3. Maybe 4. No | Most of this points to specific narrow solutions. Too early to introduce such restrictions at this stage. |
| China Telecom | (1) No  (2) No  (3) Yes  (4) Yes | Considering that data storage of RAN is possible for AI/ML data analysis, relevant research should not be excluded. And for the offline training may occur on the base station side, in this case, it’s not only temporary storage of data we are talking about. It is possible that the initial offline model training occurs in RAN, thus research can be carried out according to different use cases. The security aspects of user data storage in RAN should be coordinated with SA3. It is possible to add a statement regarding the selection of the most suitable UE based on certain criteria for AI/ML purposes. |
| Intel | 1) No  2) Yes  3) not now  4) Yes | For 1), we prefer to discuss data storage in RAN based on the use case requirement.  For 2), we agree initial offline model training should be out of RAN domain, while it can also be other entities (besides OAM). Maybe we don’t need to provide an example, “out of RAN domain” is good for RAN3.  For 3), we should discuss what data is exposed to RAN based on use case and exact information. So far, we don’t see a need to reach SA3 for RAN3 agreed use cases on LB, ES and mobility (assuming location information is not exposed to RAN).  For 4), it is important feature to reduce overload at RAN and provide AI/ML service to most needed UE for performance improvement. |
| CATT | (1) Neutral.  (2) Neutral.  (3) Not now at least.  (4) No. |  |
|  |  |  |

## 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… |
| Deutsche Telekom | Yes | We would prefer to add a note to the description of Output of Model Inference function. If this is agreeable, we can work on the text in the 2nd round of this CB. |
| Lenovo, Motorola Mobility | Yes | At least the metrics should be first discussed/defined in each use case. Then we can revisit if there is any common part can be captured in the high-level principles/functional framework. |
| Samsung | Yes | These two additional values rely on the model functionality and model output. It is fine for us to discuss in use case study. |
| Nokia | Yes | Yes, we think that accuracy and validity time should be covered on a case by case basis and where there seems to be need. It currently seems to us that validity time can be indicated implicitly by sending a new prediction when the previous one is invalid. But there could be cases/scenarios where this is not sufficient. |
| Verizon | Yes/No | Question combines multiple items for which there could be different views. As described in [7] there is value in providing validity time together with model prediction output so actor knows the duration of time for which it can reply on the output.  On the other hand, model performance metrics are specific to each model and actor is not expected to understand the metrics of every model. So their signaling is not useful. |
| Futurewei | Yes but… | In general, we agree that the details of metrics for model performance evaluation/indication do not need to be specified in the high-level principles section and can be handled at use case level. However, it is agreed in RAN3#113e and specified in TR37.817 that “An AI/ML model used in a Model Inference function has to be initially trained, validated and tested before deployment” and also “model performance feedback” is included in the RAN intelligence framework diagram. This means model performance evaluation is performed at: a) before model deployment (the outcome of model testing), and b) after the model deployment (against the labelled real-world data). Both a) and b) will produce metrics to indicate model performance. The “model performance feedback” arrow in the diagram only addresses b) but not a). To keep the diagram at high-level and simple, we suggest modifying the “model training” description in 4.2 of TR37.817 to “Model Training is a function that performs the ML model training, validation, and testing, which may generate model performance metrics as part of the model testing procedure”. |
| NEC | Yes | This is use case specific and/or solution specific. |
| China Telecom | Yes | We agree that the metrics should be discussed in each use case. |
| Intel | Ok to “validity time”  No to “accuracy/confidence level” | For validity time, we are ok to discuss per use case based, considering different meaning in each use case.  For accuracy/confidence level, it is important for other network nodes to understand whether can trust the result provided by AI/ML model, otherwise, prediction results generated by a poor accuracy model or with a poor confidence level may lead to system performance downgrade or even system break-down. Based on such accuracy/confidence level, the actor can decide whether to use AI/ML predicted action or resume legacy behavior.  Besides, accuracy level generated by Model training is important to be considered during model deployment/update, especially when we consider the inter-vendor environment between “Model Training” and “Model Inference”. |
| CATT | Yes, but… | Agree with Huawei. |

## 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? |
| Deutsche Telekom | (1) Yes  (2) Not yet | (1) We are fine adding the proposed TP into an Annex of TR 37.817 as we stated in the high-level principles we can also “think beyond” the generalized framework. Details can be further discussed.  (2) It is too early to add such a bullet point now, but we are ok to further discuss the “cost” issue and its implications in next meetings. |
| Lenovo, Motorola Mobility | No: (1) (2) | (1) Not sure if it is really necessary, it looks over specified.  (2) It is very much implementation dependent. |
| Samsung | No: (1) (2) | (1) This framework is to provide a guideline and let us have a common understanding for the terminology. It does not involve the architectural aspect. So we think maybe there is no need to map it with SA2 architectural one.  (2) It seems there is no need to set it as principle. If we need to consider cost during use case study, we can discuss it then. |
| Nokia | (1): No  (2): Yes | (1) We don’t think that such mapping is useful for our Rel. 17 study. Furthermore, introducing those entities as functions of a gNB is not granular enough to capture all possibilities.  (2) We support to introduce a bullet point regarding data collection cost. Our fear is that requests for data collection for AI/ML Training purposes may introduce a large amount of data to be sent over the interfaces. This data needs to be produced by a network node (and also transferred). A cost budget at a network node can limit/control the requested amount of data for training purposes. |
| Verizon | (1) Yes  (2) No | (1) Yes, provides a high-level understanding on how RAN AI/ML framework can be aligned to SA2 AI/ML framework  (2) Too early to do this now. Also budget/cost could mean various things that could be ambiguously interpreted. |
| Futurewei | 1. No 2. No | 1. The entities defined in SA2 are for 5GC based AI/ML operations.As the capacity and computing capability on 5GC and gNB are quite different and we don’t see the need to completely align the RAN intelligence architecture with what is specified in SA2 while leaving NG-RAN vendors to have the flexibility and decide whether to support similar functions. Such capability can be left as implementation/design issue. Alternatively, a note can be added to indicate some similar functionalities defined in SA2 for NWDAF, DCCF, ADRF on 5GC are supported in the RAN intelligence framework as well. 2. We don’t see a need to specify “cost per AI/ML measurement” as such cost may change from time to time depending on resource usage, traffic condition and other factors. |
| NEC | 1. No 2. No | 1. At this point this looks a little bit artificial alignment. Not sure what is the benefit of doing it. 2. This is one specific solution, not high level principle. |
| China Telecom | (1) Yes  (2) Yes, but | (1) The Annex related to SA2 in the AI/ML functional framework in terms of architecture are very practical.  (2) We share the same view as Deutsche Telekom. |
| Intel | 1) No  2) No | For 1), from our understanding, data services such as ADRF, DCCF is introduced as a part of service-based architecture. However, as described in the SID, Rel-17 functional framework is based on existing NG-RAN architecture, which is not suitable to define such service-based functions.  For 2), it is not clear to us how such cost budget is defined and evaluated. |
| CATT | (1) No at least now  (2) No | (2) goes too much into detail. It should be discussed per use case rather than here. |

# 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.