**3GPP TSG-RAN WG3 Meeting #111electronic R3-210980**

**25 January – 4 February 2021**

**Agenda Item: 30**

**Source: CMCC**

**Title: Summary of offline discussion on EnhDataColl**

**Document for: Discussion and Decision**

# 1 Introduction

**CB: # 17\_Basket\_AI**

**- resolution of FFSs**

**- high-level framework for continuation of SI**

**- new use case: energy saving? (lower prio discussion?)**

**- If consensus, capture any general principles / descriptions / open issues / WA / agreements for upcoming RAN3 work (to be captured in Chair’s Notes)**

**- suggest to avoid discussing details (CRs,TPs, etc.) unless there is full agreement**

(CMCC - moderator)

Summary of offline disc R3-210980

We intend to achieve some high-level agreements during the first phase discussion and proceed with the TPs if possible in the second phase.

# 2 For the Chairman’s Notes

**To be added after email discussion.**

# 3 Discussion

## 3.1 TR 37.817 v0.1.0

The TR [1] is updated based on the agreements on RAN2#110 E-meeting. It has been checked over the email reflector after RAN3 #110e meeting and resubmitted to RAN3 #111e for agreement.

**Proposal 1. Agree the TR 37.817 v0.1.0**

## 3.2 High-level AI framework

Following are open issues left for AI framework:

*Editor's Note: the details for the framework below is FFS including whether Actor and Subject of action should be in one box or separate, whether feedback from action to Model training host is needed, the name in each box is from functionality or from processing point of view, the feedback from Subject of action to the Data sources is Performance feedback or Model performance feedback and other possible refinement.*

Paper [2][3][4] address on these open issues.

### **3.2.1 AI framework from functionality or from processing point of view**

In R3-210917, two alternatives for illustration of the AI functional framework are discussed:



Figure 1 Alternative 1: AI framework from functionality point of view (currently captured in the TR 37.817)



Figure 2 Alternative 2: AI framework from processing point of view

After some comparison, it is found that the two alternatives do not differ two much. It is proposed that if we cannot reach consensus at this stage, we could keep the current alternative as in the TR and refine it at later stage when we found it is not suitable for use case and solution description.

On the other hand, it is pointed out in R3-210617 that, currently in Figure1, each box represents one processing host to enable AI functionality. However, for one AI functionality in one use case, multiple processing hosts may be used. On the other hand, there is a possibility that more than one AI functionality can be supported by the same one processing host. Therefore, it is preferred for each box to represent one processing action to enable AI functionality including data collection, model training, model inference and action.

Companies are invited to provide views on whether to keep the current alternative as in the TR or choose alternative 2:

**Q1: Do you agree to keep the current alternative as in the TR or choose alternative 2?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Reasons/Comments/Suggestions** |
| ZTE | Prefer AI framework from processing point of view.  (Alternative 2 is partly OK) | The framework we proposed as follow:    The framework aims to show the entire the ML operation process for RAN intelligence. The framework should be separated from the RAN architecture (including RAN logical node). We prefer the framework comprising four boxes including data collection, model inference, model training and action.  The difference between alternative 2 and framework:  - Performance feedback from action to model training.  - Clarify that model training includes the online training. |
| Deutsche Telekom | We are ok to go with alternative 2 which focuses on functional blocks only (i.e., leaving hosts out of the figure). | With alternative 2 we can address a pure functional AI framework without addressing any deployment aspects. Based on that functional framework, use case specific deployments can be considered in a later phase of the study.  Following proposals for updates of Figure 2 (except of those discussed under following sections):   * Change “Data collection” to “Data collection & preparation” to make clear that this function is not simply collecting data, but also separate it and possibly pre-process it for different purposes. * Change “Model training” to “Model training (offline/online)” to make clear that this function may address both variants.  Note: In a use case specific deployment the function may be split into an offline and an online part which may be placed in different network nodes. |
| China Unicom | We are OK with ZTE’s proposal. | We are fine with the processing model for RAN side. The modifications in ZTE’s proposal are essential for Model training. |
| Nokia | We think both options are OK but | Alternative 2, representing the processing action is more simple since it does not address deployment aspects, e.g., related to where inference or training will be hosted.  We agree with Deutsche Telecom that we need to update Data Collection to “Data Collection & Preparation” to capture not only the process of collecting data, but also the process of preparing/pre-processing the input data for training and inference. |
| vivo | We are fine with Alt2 | Alternative 2 is much focused on AI functional description. We think whether to add “including online training” as proposed by ZTE is just adding additional description for the “Model training”, we are open to consider such addition. |
| Huawei | Fine with Alt 2 | Similar view as mentioned above, each block in alt2 reflects more as functionality, we are also ok with DT’s suggestion. |
| NEC | Both alternatives need some modifications | We think that Actor and Subject of Action should be kept separate.  Do we need multiple Subjects of Action?  We support proposal to have feedback from Subject of Action to Model Training.  How to capture that output from one model could be input to another model? |
| Qualcomm | Slightly prefer Alt 2 | Both alternatives are fine. Using a single box for action is clear. |
| Samsung | Both alternatives are fine. | Alternative 1 is made up of the functionality blocks (not deployment aspects) and relevant interaction between each functionality block. It clearly shows the function & the input/output of each component and the working relationship among all participators, which is beneficial for exploring the network node functionality and interface impact.  Alternative 2 shows the process of AI/ML operation, the functionality of each step and the related input/output of each step, which is beneficial for exploring the lifecycle management of the AI function implementation. |

### **3.2.2 Whether Actor and Subject of action should be in one box or separate?**

In R3-210917, it is observed that for the purpose of easy identifying the signalling exchange between network nodes, the separation of actor and subject of action seems to be beneficial.

In R3-210785, it is pointed that one box is enough to reflect the function of action. In this framework, the process of the AI should be clearly defined, and the place where model training, model inference, data collection and action (involve actor and subject of action) needs to be discussed case by case.

In R3-210617, it is proposed that to make the Functional Framework simple and straightforward, the separate “Actor” and “Subject of action” boxes can be merged into one block (e.g. Action). The “Action” box does not restrict that only one node or interface is involved for one action, if clear explanation is necessary, one note can be added that “one or more Subjects of Action(s) may act over at least one interface”.

Companies are invited to provide views on whether actor and subject of action should be in one box or separate:

**Q2: Whether actor and subject of action should be in one box or separate?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Reasons/Comments/Suggestions** |
| ZTE | Yes | The framework aims to show the entire the ML operation process for RAN intelligence. The framework should be separated from the RAN architecture (including RAN logical node).Model inference executes the trained model to get the output based on the inference data. Then, specific policy needs to be generated based on the output. One box is enough to reflect the function of action. |
| Deutsche Telekom | Separate boxes preferred. | With the separation it is more clear that the model interference does not trigger a direct action, but provides information to a function called “actor” which is responsible for triggering actions in different other functions which may be placed in the same node or other nodes of the network. |
| China Unicom | Yes | If considering the functional framework, one box is simple for description. But If considering network deployment, it is uncertain to deploy in one node or separate nodes. |
| Nokia | Separate boxes | Since actor and subject of action may be located in different entities, it is good to be able to capture this in the framework. This separation might also reveal standardization impacts over the interfaces depending on the use case. |
| Vivo | Separate boxes | For more coherent picture of actor and subject, we prefer to have separate boxes. |
| Huawei | One box | From functionality point of view, one box should be enough, whether the action would be taken by two separate entities, this would depend on the concrete use case where whether two entities should be involved or not. |
| NEC | Separate boxes | Actor and subject of action could be located in different RAN nodes or functional entities. Having them in separate boxes is helpful to study standardization impacts on interfaces. |
| Qualcomm | One box | This figure just shows the big function split and workflow. One box for action is clear. This does not prevent from supporting action in multiple network nodes. |
| Samsung | Separate boxes | The separation helps to investigate standardization impacts over the interfaces depending on the use case. |

### **3.2.3 Whether feedback from action to data sources is performance feedback or model performance feedback**

Both R3-210917 and R3-210785 think the name “Model performance feedback” is appropriate. Since ML inference is a process of using a trained ML model to make a prediction or guide the decision based on collected inference data and ML model. The output can be feedback to the model training host to verify the performance of the ML model and in turn help the model training host to improve or re-select the ML model.

Companies are invited to provide views on whether feedback from action to data sources is performance feedback or model performance feedback:

**Q3: Whether feedback from action to data sources is performance feedback or model performance feedback?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Reasons/Comments/Suggestions** |
| ZTE | Feedback from action to data sources is performance feedback.  Feedback from inference to training is model performance feedback. | Feedback from action to data sources is performance feedback.  For the feedback between Model inference and Model training, since Model inference is one component which execute the trained model to get the output based on the inference data, this performance reflects the ML model performance is good or not. If this model performance feedback is not good, ML model needs to be re-selected or re-trained in the Model training component. So feedback between Model training and Model inference is called “Model performance feedback”. Taking AI-based energy saving as an example, the output of the model inference may be the predicted load of the cell.  For the feedback from Action to data collection, this feedback reflects the network performance after the Action. After the policy is adopted , the network performance may be optimized/maintained/degraded. Taking AI-based energy saving as an example, the feedback from Action is the KPI of the network. |
| Deutsche Telekom | Use of “Model performance feedback” is supported. | This term was already proposed in our contribution to the last meeting (see R3-206197). |
| China Unicom | Prefer performance feedback | The data types for performance feedback are different from the data for model performance feedback. It is preferred to have two definitions and performance feedback is the feedback from action to data sources/data collection. |
| Nokia | Performance Feedback | Model performance feedback relates to the accuracy of the model or prediction quality. Feedback from Action to Data Source is related to the impact of the action taken given the model output and is related to performance. |
| vivo | Performance Feedback | Performance feedback to allow assessment of model performance for further training enhancement. |
| Huawei | Performance Feedback | Model performance feedback, our understanding, is to evaluate the trained model accuracy, while the performance feedback is to evaluate the practical effect of the action from the trained model. |
| Qualcomm | Performance feedback | Agree with ZTE |
| Samsung | Performance feedback | The feedback from action to data sources is to transfer the network performance after applying AI-generated or AI-assisted policy, such as QoS or other KPI parameters. The network performance data can be used as input to training and/or inference with the aim as 1) feedback the model decision impact for RAN so as to improve the model efficiency; 2) describe the current network status.  For the feedback from inference host to training host is the model evaluation value to evaluate whether model works well (e.g. the accuracy to evaluate a prediction model), which is more proper to be named as model performance feedback.  Thus, we prefer to keep performance feedback to name the feedback from action to data sources. |

### **3.2.4 Whether feedback from action to model training host is needed**

Some company deems that performance feedback from action to data source is enough [2], but others find that the feedback from action to model training is needed for re-training or reinforcement learning [3].

Companies are invited to provide views on whether feedback from action to model training host is needed:

**Q4: Whether feedback from action to model training host is needed?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Reasons/Comments/Suggestions** |
| ZTE | Yes | Model training is one component that train the ML model offline or online according to different use case, so re-training or updating model is also one part of model training. Model training may be triggered to retrain the ML model when the model performance feedback is not good. Similarly, if the network performance feedback from Action is not good, Model training should also be triggered to retrain the ML model or update the ML model.  On the other hand, reinforcement learning, as one of the ML techniques (online training), needs to interact with the network environment during training phase. During training phase, the agent in reinforcement learning needs to get the reward from the environment through performance feedback from action. If the framework should support the reinforcement learning, the performance feedback to the model training is essential. |
| Deutsche Telekom | Yes, but … | Taking alternative 2 as baseline for functional framework where “Model training” should cover both online and offline training, the direct feedback loop is only required for online training (optional link). This can be clarified by using e.g. a dashed line in the figure and additional text for explanation.  Re-training/reinforcement learning aspects have to be considered in use case specific deployment scenarios where both training schemes may be in different hosts/nodes and therefore the data flows may be also split. |
| China Unicom | Yes | We agree to implement online and offline training, and the feedback from action to model training host is necessary |
| Nokia | Yes | Feedback from Action to Model Training can be useful for online training and reinforcement learning. |
| vivo | Yes | Agree with China Unicom. This is necessary for online training and ML enhancement. |
| Huawei | Yes | Similar view as Nokia. |
| NEC | Yes | This is needed for online/reinforcement training. |
| Qualcomm | Yes | It is useful for online training and reinforcement training. |
| Samsung | NO | There are multiple factors to affect the network performance, such as channel condition, generated policy, equipment issues etc. If the network performance is bad, the analysis is required to be done to find out the reason firstly. When the problem is coming from AI model based on analysis, the model needs to trigger the retraining procedure, otherwise, the model-retraining should not be triggered. So the retraining procedure cannot be triggered directly by the network performance.  For the reinforcement learning, the performance data can be one of the training inputs. For the current framework, the performance data can be obtained by training host from data sources as training data, which is more reasonable for the role of performance data for reinforcement learning.  Thus, we prefer the framework without feedback from action to model training host. |

### **3.2.5 Other open issues for AI framework**

In R3-210617, some other open issues are discussed and following proposals are proposed:

**Proposal 3: RAN3 supports the case that one ML model demands input from other ML models.**

**Proposal 4: For the sake of discussion, RAN3 further distinguishes the decision-oriented ML model and the prediction-oriented ML model when it comes to ML model feedback provision and ML model retraining/update.**

**Proposal 5: The data collection shall provide the training data according to the demand of model training regarding what/when/how to provide. It is regarded as model performance feedback when the training data provision is triggered due to bad model performance.**

**Proposal 6: The data collection also collects the output from the model inference.**

Since above issues have not been touched in last RAN3 meeting, companies are invited to provide views on whether to discuss these open issues in this offline discussion:

**Q5: Whether to discuss above open issues proposed in R3-210617 in this offline discussion?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Reasons/Comments/Suggestions** |
| ZTE |  | P3, P4 may discuss case by case later..  For P5, we think there is no need to add the data provision policy from model training to data collection into the current AI framework. The framework aims to show the entire the ML operation process for RAN intelligence rather than signaling design  For P6, not needed, model training can get the prediction result through the model performance feedback from model inference to model training. |
| Deutsche Telekom | No | Due to 0 TUs in this meeting we should focus on items in Sec. 3.2.1 – 3.2.4 only. |
| Nokia | No | We agree with Deutsche Telecom. |
| vvio | No |  |
| Huawei | Pending on the available TU | Given there is TU available, we could discuss, but the proposals themselves are not very clear, e.g. for P3, is it referring to federal learning? for P4, I suppose it is related Q6? |
| Qualcomm | No | Agree with DT |
| Samsung |  | Same view with DT. Prefer to have the discussion at next meeting with allocated TU. |

## 3.3 Use case

### **3.3.1 Use case priority**

In R3-210073, it is observed that most interested use cases are: energy saving, traffic/load prediction, trajectory prediction, traffic steering and load balancing. Moreover, some of them can be used as a tool box by other use cases, e.g., load prediction, UE trajectory prediction. Therefore, it is proposed to study the below use cases in R17 with high priority:

1. Tool box use cases:

- traffic/load prediction

- trajectory prediction

1. System level use cases:

- energy saving

Companies are invited to provide views on whether to study above use cases in R17 with high priority:

**Q6: Whether to study above use cases in R17 with high priority?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Reasons/Comments/Suggestions** |
| ZTE | Yes | The most interested use cases are: energy saving, traffic/load prediction, trajectory prediction, traffic steering and load balancing.  Some of them can be regarded as system level use case. Take energy saving as an example, model training can be deployed outside RAN, e.g., OAM system, the whole system solution including three main functions, AI based scenario classification, AI based Load prediction, and AI based ES performance analysis.  While some of them can be used as a tool box by other use cases, e.g., load prediction, UE trajectory prediction. For example, load prediction can contribute to use cases like energy saving, load balancing, mobility optimization. UE trajectory prediction can also contribute to use cases like mobility optimization, energy saving. For those tool box use cases which can be achieved purely on the RAN side for RAN purpose. |
| Deutsche Telekom | Yes, but … | We are in principle ok to start with those use cases, but nevertheless, this should not prevent other use cases to be considered within the SI phase (contribution-driven). |
| China Unicom | Yes | The above three use cases are good start for R17, but we should not prevent other use case in the future discussion. |
| Nokia | Not quite | Energy saving is one of the approved use cases to start with and we support it. However, traffic/load prediction and trajectory prediction are not really use cases in our view but are, rather, enablers that could be used by other use cases.  We support to start by the approved use cases of energy saving, load balancing, traffic steering/mobility optimization. To allow some diversity in the cases we consider, we also support use cases for optimization of lower layer parameters. |
| vivo | Yes | We agree to support all above use cases with high priority. From our understanding these are basic scenarios subject to AI. |
| Huawei | Yes for (2) | Similar view as Nokia, for energy saving for sure yes. For (1), we think they are just intermediate steps which would be applied to different use cases, such energy saving, load balancing, mobility optimization, etc. |
| NEC | Yes for energy saving | We support to start from use cases selected in 110e but allow to consider other use cases case by case. |
| Qualcomm | Yes | Energy saving has been agreed.  Tool box use cases can provide input and support for other use cases. So, 1) should have higher value than the system level use cases. |
| Samsung | Yes for (2) | Prefer the agreed use case in 110e meeting to be the start point. And discuss the other use case at next meeting with allocated TU. |

### **3.3.2 Use case description**

R3-210918, AI based Energy Saving is used as the starting point and intend to work out an example on how to describe the use case and solutions. Companies are invited to provide views on the description if any:

**Q7: Do you agree to use Energy Saving as the starting point and what’s your comments about the description in R3-210918 if any?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Reasons/Comments/Suggestions** |
| ZTE | The detail use case description can be further discussed in the next meeting. | As we analyzed above, there are three main functions for AI based energy saving (AI based scenario classification, AI based Load prediction, and AI based ES performance analysis), for each function, how to deploy the AI framework into the current system architecture, and the detail inputs and outputs are different, which needs to be further discussed. |
| Deutsche Telekom | No | Due to 0 TUs in this meeting we should focus on items in Sec. 3.2.1 – 3.2.4 for the functional framework only. Going into details of a use case description does not make sense without a stable definition of a functional framework that can be used as a basis. |
| China Unicom | The detail description can be further discussed. | AI based energy saving is a good use case, but the detail of the description of that could be discussed in next meeting. Before we discuss the use case, the framework needs to be defined at the first stage. |
| Nokia | No | We should first focus on the ML framework and then go into details of specific use cases. We cannot study use cases before having defined a solid functional framework. |
| vivo |  | Energy saving should one of the use cases to start with, but not the only one. We think description of this use case should be further considered. |
| Huawei | Yes | Anyway we should focus on some use cases, now we saw more than ten use cases, if we have no focus, it is difficult to converge I am afraid. And, energy saving is one of the most useful features contributing OPEX reduction which is so important to operators. |
| NEC |  | Detailed use cases could be discussed in the next meeting. |
| Qualcomm |  | Agree with ZTE. The description can be discussed in next meeting based on contributions. |
| Samsung |  | Detailed use case can be discussed in the next meeting. |

# 4 Conclusion, Recommendations

To be edited, if needed**.**

# 5 Reference

1. R3-210935, TR 37.817 v0.1.0 (CMCC)
2. R3-210917, Further discussion on high-level AI framework (CMCC)
3. R3-210785, High-level framework and definition for AI RAN (ZTE Corporation)
4. R3-210617, Open issues of framework for AI (Lenovo, Motorola Mobility)
5. R3-210918, AI based Energy Saving (CMCC)
6. R3-210073, Use cases for AI study (ZTE Corporation)