Pro3GPP TSG-RAN WG3 Meeting #113-e R3-214222

E-meeting, 16– 26 August, 2021

**Agenda item: 18.4.2**

**Source: Nokia (moderator)**

**Title: CB: # AIRAN4\_LBSolution - Summary of email discussion**

**Document for: Approval**

# 1 Introduction

**CB: # AIRAN4\_LBSolution**

**- Discuss the solution, input/output, standard impacts on the Load Balancing**

**- Merging any agreement parts; provide TP if agreeable**

**- Capture agreements and open issues**

(Nok - moderator)

Summary of offline disc in [R3-214222](file:///D:\Meetings\RAN3%23113\CB\Inbox\R3-214222.zip)

The deadline for the first phase of the email discussion is Friday 6 am UTC.

# 2 For the Chairman’s Notes

For the second round of discussions we propose to check moderator’s proposals from the first round and additionally a TP on the AI/ML load balancing solution. Please provide your further comments by Tuesday 24th, at 8am UTC.

**Proposal 1: The following solutions for the Load Balancing use case included in the TR:**

* **AI/ML Training is located in the OAM and AI/ML Inference is located in the gNB.**
* **AI/ML Training and AI/ML inference are located in the gNB (e.g., for online training).**

**Proposal 2: In CU-DU split, the following solutions for the Load Balancing use case are included in the TR:**

* **AI/ML Training is located in the OAM and AI/ML Inference is located in the gNB-CU.**
* **AI/ML Training and Inference are located in the gNB-CU.**
* **Other solutions are FFS.**

**Proposal 3: We do not need categorization of the AI/ML Load Balancing solution in AI/ML assisted or AI/ML generated.**

**Proposal 4: gNBs may signal load predictions to each other over Xn interface. It is FFS whether a gNB-CU may request load predictions from a gNB-DU it manages.**

**Proposal 5: It is FFS whether prediction accuracy, validity time, deadline need to be signaled in the input/output of AI/ML inference. More clarifications/definitions are needed for their definition and usability.**

**Proposal 6: If existing UE measurements are needed by a gNB for the load balancing use case, RAN3 shall consider the MDT framework as a baseline. FFS whether new RRC signaling is needed.**

**Proposal 7: Agree the TP on AI/ML Load Balancing use case.**

# 3 Background

In RAN3 #112-e, the description for the Load Balancing use case was agreed by RAN3. According to the use case description in TR 37.817, load balancing optimization is challenging for several reasons:

* Currently the load balancing decisions relying on the current/past-state cell load status are insufficient. The traffic load and resource status of the network changes rapidly, especially in the scenarios with high-mobility and large number of connections, which may lead to ping-pong handover between different cells, cell overload and degradation of user service quality.
* It is difficult to guarantee the overall network and service performance when performing load balancing. For the load balancing, the UEs in the congested cell may be offloaded to the target cell, by means of handover procedure or adapting handover configuration. For example, if the UEs with time-varying traffic load are offloaded to the target cell, the target cell may be overloaded with new-arrival heavy traffic. It is difficult to determine whether the service performance after the offloading action meets the desired targets.

Through AI/ML, the load balancing performance can be improved.

In the first round of this CB, we discuss different solutions on the location of the AI/ML functionality for the load balancing use case, whether the load balancing solution should be categorized into AI/ML generated or AI/ML assisted, the exchange of load predictions between gNBs or between a gNB-CU and a gNB-DU it manages, and the possibility for a gNB to request assistance information for load balancing purposes. In the second round of the email discussion, we will attempt to identify some agreeable inputs and outputs to the Model Inference function once the different solutions for placing the AI/ML functionality have been agreed.

# 4 Discussion - 1st Round

## Location of the AI/ML Functionality

Different options have been discussed regarding where the AI/ML functionality may be placed. Contribution 3296 supports both local and centralized solutions. Contributions 3420, 3714, and 4080 support AI/ML inference in the RAN. The possible options proposed by different companies can be categorized as follows: a) AI/ML Training and AI/ML Inference may be located in OAM (3470, 3758, 4124), b) AI/ML Training may be located in OAM and AI/ML Inference may be located in the gNB (3470, 3758, 4124), c) AI/ML Training and Inference may both be located in the gNB (3470, 3893).

**Q4.1-1 Companies are invited to discuss which of the following solutions they prefer. More than 1 solution is possible.**

|  |  |  |
| --- | --- | --- |
| **Company** | **Solution 1-1.a), 1-1.b), 1-1.c)**  **(More than one selection may be chosen)** | **Comment** |
| Huawei | b) is preferred | As discussed in the paper on general framework, we think training is in general an offline action which requires storing large amount of data, this is against the general principle that RAN is not a place for storage. |
| Ericsson | b), c) | Solution a) is out of scope of the SI. That can be deduced from the SID (RP-201629), stating that the objective is:  “Study high level principles for RAN intelligence enabled by AI”  With this we are not saying that these solutions are not appropriate, but they are not the subject of this SI  Solution b) is perhaps the solution range we should give highest priority. Namely, the RAN is given a trained model and we would study procedures to support inference at the RAN  Solution c) is perhaps the second highest priority type of solution, where training and inference occurs at the RAN |
| Nokia | b), c) | Solution a) is not in the scope of SI since all the AI/ML functionality is in the OAM. |
| Intel | Solution 1-1.a), 1-1.b), 1-1.c) with comment | As discussed in CB: #AIRAN2, we need further check with SA5 on the feasibility of solution b).  However, solution a) is under scope of SA5, which is already defined. |
| vivo | b) | option a is under SA5 scope.  For option c, numerous message exchange between RAN nodes is essential as one single RAN node cannot acquire enough information for model training. |
| Samsung | b) | As training needs a large volume of data storage and computation power, OAM is a suitable location for training. The model inference requirements of data storage and computation power are not quite high. So to realize the real time prediction to provide the latest output to optimize RAN, gNB is a suitable location for model inference. |
| Lenovo and Motorola Mobility | b) c) | AI/ML Training for load balancing can be performed online in gNB or offline in OAM.  AI/ML Inference for load balancing can be located in gNB. |
| CMCC | 1. b) c) | All the three options are feasible. b) and c) are more in the responsibility of RAN3, which are high priority. |
| KDDI | b), c) | Option a) is under SA5 scope. Considering computation power for AI/ML, b) is the first priority and c) is the second priority. |
| ZTE | a), b), c) | For the solution a), there is no standard impact in RAN3 if Model training and Model inference are both located in the OAM.  For AI-based load prediction, inference stage also can be located in the NG-RAN node in case that the optimization decision should be made according to the short-term predicted load. So we can focus on the solution b) and solution c) in RAN3. |
| Deutsche Telekom | b) | a) is out of scope of this SI as it is a pure OAM case and can be considered by SA5 in their MDA work.  c) is not preferred as initial (offline) ML training should not happen from our perspective in the RAN domain (e.g. due to data storage and processing power required). |

**Summary:**

3 companies support option a) AI/ML Training and AI/ML Inference may be located in OAM

11 companies support option b) AI/ML Training may be located in OAM and AI/ML Inference may be located in the gNB

7 companies support option c) AI/ML Training and Inference may both be located in the gNB

With the understanding that option c) includes online training and option a) is not in the scope of the SI, we make the following proposal:

**Proposal: The following solutions for the Load Balancing use case are considered:**

* **AI/ML Training is located in the OAM and AI/ML Inference is located in the gNB.**
* **AI/ML Training and AI/ML inference are located in the gNB (e.g., for online training).**

In case of split architectures, the following possibilities on placement of the AI/ML functionality have been discussed by different companies: a) AI/ML Training in the OAM and AI/ML Inference in the gNB-CU (3725, 3758), b) AI/ML Training in the OAM and AI/ML inference in the gNB-DU (3758), and c) AI/ML Training and Inference in the gNB-CU (3725, 3893).

**Q4.1-2 Companies are invited to discuss which of the following solutions they prefer. More than 1 solution is possible**

|  |  |  |
| --- | --- | --- |
| **Company** | **Solution 1-2.a), 1-2.b), 1-2.c)**  **(More than one selection may be chosen)** | **Comment** |
| Huawei | 1. Is preferred | For training, as commented above, it should be located in OAM; for inference, it should be located in gNB. As to gNB-CU/DU split architecture, we think the final load balancing decision is anyway made by gNB-CU, since gNB-CU knows the whole situation more than gNB-DU does. |
| Ericsson | a), b), c) | We need to first converge on more details on the solution. In our proposal in R3-213420 it is the RAN node that derives predicted load metrics. Load metrics are in part produced by the gNB-DU and signalled to the gNB-CU. We need more discussions to determine whether prediction is done only by the gNB-CU or only by the gNB-DU or by both |
| Nokia | a), and c) | It is a bit difficult to understand how b) would work having training in OAM and inference in the DU. In our view, we should focus on the more straightforward solutions at first to avoid overload of the SI. |
| Intel | 1-2.a), 1-2.c) and training at CU/inference at DU | We understand the intention of solution b). However, we think it would be good to consider mode training at gNB-CU, and inference at gNB-DU (i.e. combining b) and c)). The outcome of Model inference at gNB-DU can also be beneficial to other use cases as well, for example, scheduling optimization based on load prediction. Compared with solution c), this can reduce latency for exchanging predicted results which will be used for real-time use cases. Hence, we think there might be a benefit to keep model inference at gNB-DU. |
| vivo | a) | CU is the suitable node for AI model to reside as CU has the over view of the resource status and some input of the model may be collected at CU, e.g. the RRC connection number. |
| Samsung | a), b) | CU is a suitable node for inference as it can collect the resource status from neighbours and the load balancing decision is made in CU.  For DU load prediction, the model inference may locate at DU. As the model inference needs the computation power, resource status of all DU predicted by CU may bring computation burden to CU. |
| Lenovo and Motorola Mobility | a), c) | AI/ML Training and Inference for load balancing can be performed in gNB. For CU-DU split architecture, CU is more suitable than DU to train ML model or perform inference, since CU has an overall view for the resource status of the related DUs via the existing F1/E1 interface, and CU has more computation resource for load prediction. |
| CMCC | 1. b) c) | In CU-DU architecture, load could be calculated in CU and DU depending on the load metrcs, e.g., PRB usage are in DU, RRC connections are in CU, so similar view as Ericsson, We need more discussions to determine whether prediction is done only by the gNB-CU or only by the gNB-DU or by both |
| KDDI | a), b), c) | Option b) is needed to discuss the following points.  1. How to work having training in OAM and inference in the DU (discussed in SA5?)  2. Load metrics related to DU (ex. Radio resources) |
| ZTE | a), c) | For CU/DU split architecture, what we proposed in 3758 is AI/ML training in the OAM and AI/ML inference in the gNB-CU, or AI/ML training and inference both in gNB-CU, or AI/ML training in gNB-CU and AI/ML inference in gNB-DU. In short, we think gNB-CU can be responsible for the ML-training because CU may have better computation capability than DU, and then gNB-DU can be responsible for the ML inference. For model inference at CU or DU, the corresponding input/output data will be different. |
| Deutsche Telekom | a) | Same view as Huawei. |

**Summary:**

11 companies support option a) AI/ML Training in the OAM and AI/ML Inference in the gNB-CU

4 companies support option b) AI/ML Training in the OAM and AI/ML inference in the gNB-DU

7 companies support option c) AI/ML Training and Inference in the gNB-CU

**Proposal: In CU-DU split, the following solutions for the Load Balancing use case are considered:**

* **AI/ML Training is located in the OAM and AI/ML Inference is located in the gNB-CU.**
* **AI/ML Training and Inference are located in the gNB-CU.**
* **Other solutions are FFS.**

## AI/ML-assisted Load Balancing versus AI/ML-generated Load Balancing

In 3714, it is proposed that AI/ML functionality should be categorized into Type 1: AI/ML-assisted Load Balancing where Load Balancing strategy is calculated using the conventional Load Balancing method that utilizes load predictions and b) Type 2: AI/ML generated Load Balancing where the Load Balancing strategy is calculated by using the current/historical resource status.

**Q4.2.1 Companies are invited to provide their views on the categorization of the Load Balancing solution.**

|  |  |  |
| --- | --- | --- |
| **Company** | **AI/ML assisted, AI/ML generated, both types, no need for categorization** | **Comment** |
| Huawei | Either way could work | We think this is an implementation choice. Maybe we should first to investigate if there are any differences between the two solutions, concerning the spec impacts. In our limited understanding, seems no big differences, from input and output perspective. |
| Ericsson | None | We do not understand why this distinction is needed. In our functional framework we have clear behaviours and roles for different functions. We are studying solutions where an Actor function takes actions based on a received Model Inference function output. We do not specify whether these actions are based only on the Model Inference output, or on the output plus historical data, etc. Namely the Actor “may” take the Model Inference output into account to perform its actions and it is up to the Actor implementation to decide how to ultimately take the action. Hence, the classification seems not to be of relevance for the work RAN3 is carrying out. |
| Nokia | No need for categorization | These categories just depend on how different ML Algorithms operate to produce their output. It should not matter for the purpose of the SI. |
| Intel | no need for categorization | We share the same view as HW that two solutions may have similar signalling impact. We may not need to differentiate which one is used. |
| Vivo | no need for categorization | Type2 includes type1 as the load prediction is the key issue of the AI based LB. Proponent can clarify the necessary, e.g. the procedure or message exchange are different. |
| Samsung | Yes | This classification is from AI/ML functionality aspect instead of algorithm aspect. For these two types, the AI functionality and corresponding input/output are not same. The classification can help to sort out the standard impact for different AI functionalities. |
| Lenovo and Motorola Mobility |  | Since we are discussing solutions for load balancing, it’s of course important to understand what the purpose of the AI model is, meaning the output of the AI model. Type 1 is providing predicted input for a load balancing decision. Type 2 is generating the load balancing decision. We will not discuss the exact algorithm, but classifying the solutions into type 1 and 2 sounds reasonable and will facilitate the discussion. On the other hand, whether we need to explicitly define these 2 types of models or just explain in the solution description whether this solution is to generate a load balancing decision or to provide input for a load balancing decision, we are open. |
| CMCC | Seems reasonble | For AI/ML-assisted Load Balancing, the output could be predicted load, while for AI/ML-generated Load Balancing, the output of model inference could be HO decisions. The signalling procedures and spec impact nay be different. We could not illustrate the two directions in the TR. |
| KDDI | Reasonable | Classifying the solutions seems to be reasonable for facilitating the discussion. If signalling procedures and spec impact between two types are actually different, this classifying is needed to progress SI. |
| ZTE | No need for categorization | We prefer not to classify the solutions. Type1 is to do load prediction, while type2 is to generate the LB decision directly by ML model. Two kinds of the solution may have similar signalling impact. |
| Deutsche Telekom | Not needed | Both solutions are possible (probably also more?), but we don’t need that strict classification here, as this is more related to solution implementation. |

**Summary:**

On the question whether the AI/ML Load Balancing solution needs to be categorized in AI/ML assisted or AI/ML generated:

6 companies think that there is no need for categorization

3 companies think that there is need for categorization

2 companies are open to either way

**Proposal: We do not need categorization of the AI/ML Load Balancing solution in AI/ML assisted or AI/ML generated.**

## Exchange of Load Predictions between Neighbours and between a gNB-CU and a gNB-DU in Split Architectures

Contributions 3420, 3714, 3725, 3758, 3893 propose that a gNB should be able to request a neighbouring gNB to start calculation and reporting of load predictions. In case of split architectures, contributions 3420, 3714, 3893 propose that a gNB-CU can request load predictions from a gNB-DU connected to it.

**Q4.3-1 Companies are asked to provide their views whether gNBs may signal load predictions to each other as well as whether in split architectures load predictions can be sent between a gNB-CU and a gNB-DU it manages.**

|  |  |  |
| --- | --- | --- |
| **Company** | **3-1.a) Introduce load prediction reporting between gNBs?**  **3-1.b) Introduce load prediction reporting between a gNB-CU and a gNB-DU its manages?** | **Comment** |
| Huawei | For 3-1.a): OK  For 3-1.b): not sure | We are not sure why gNB-CU could not predict the load for each gNB-DU connecting to it, assuming that gNB-CU anyway will know the real time load situation of each gNB-DU. Besides, since a gNB-CU is able to collect information from its DU cells and neighbour cells, the load predictions from a gNB-CU is more accurate. |
| Ericsson | 3-1.a) OK  3-1.b): needs clarification | We understand that 3-1.b) is about the gNB-DU predicting load for its own cells and signalling this to the gNB-CU. We would like to keep this point FFS and investigate more about it |
| Nokia | 3-1.a) OK  3-1.b) OK | Load predictions received at a gNB-CU from neighbour gNBs (gNB-CUs) as well as from a gNB-DU it manages can provide the gNB-CU with needed input for the load balancing use case. For 3-1.b), a gNB-CU could also ask for resource status information and calculate load predictions itself instead of asking the DU directly for load predictions. This would be another alternative.  To clarify E///’s understanding, 3-1.b) is indeed about gNB-DU calculating load predictions over its own cells and sending those to its gNB-CU. It is also OK to leave this point FFS. |
| Intel | 3-1.a) Agree  3-1.b) Agree | If model inference is located at gNB-DU, the load prediction result may need to report to gNB-CU so that a proper load balancing at gNB-DU can be performed.  Additionally, this prediction information can also considered to be further used by other AI/ML models in gNB-CU as input (as we discussed in CB #AIRAN2). |
| vivo | 3-1.a): OK  3-1.b): not sure | CU is responsible for the model inference. For option b, the CU can request for the input from DU to derive the load prediction. |
| Samsung | 3-1.a): OK  3-1.b): OK | As the model inference needs the computation power, resource status of all DU predicted by CU may bring computation burden to CU. |
| Lenovo and Motorola Mobility | 3-1.a) ok | For a, agree with load prediction reporting between gNBs.  For b, in case of split architectures, we prefer it is gNB-CU that predicts load information based on the inputs e.g. current/historical resource status from the related gNB-DUs, rather than gNB-CU requests load predictions from gNB-DU. |
| CMCC | Ok | But the second one in CU-DU architecture, depends on the answers to the question **Q4.1-2** |
| KDDI | 3-1.a) OK  3-1.b) OK |  |
| ZTE | Both OK | The predicted load information is transferred to the neighbouring NG-RAN nodes or gNB-CU via Resource Status Report procedure which can be used for subsequent optimization.  For ML inference in CU, calculating predicted load resides in the CU so there is no need to send load prediction from DU to CU, but before prediction, CU needs to request inputs data from DU (e.g., historical load status).  For ML inference in DU, calculating predicted load resides in the DU, it is better that CU requests load prediction from DUs for subsequent optimization. |
| Deutsche Telekom | 3-1a) Ok  3-2b) Not sure | W.r.t. b) CU should be able to derive load prediction based on available info from its DUs. |

**Summary:**

11 companies agree that gNBs may signal load predictions to each other over Xn interface.

In case of CU-DU split:

* 6 companies agree that a gNB-CU may request load predictions from a gNB-DU it manages
* 3 companies are not sure
* 1 company supports that gNB-CU predicts load information based on inputs e.g. current/historical resource status from the related gNB-DUs instead
* 1 company needs clarification

**Proposal: gNBs may signal load predictions to each other over Xn interface. It is FFS whether a gNB-CU may request load predictions from a gNB-DU it manages.**

It was also discussed by some companies that a node requesting load predictions from another node should include in the request also an accuracy the requested load predictions need to satisfy (e.g., a prediction accuracy (3420, 3725, 3893)), a time for which the predictions must be valid (validity time (3420)), and a deadline by which predictions must be provided (3893).

**Q4.3-2 In case an answer to Q4.3-1 is yes, companies are invited to provide their views on the need to include in the request for load predictions a) a prediction accuracy, b) a validity time and c) a deadline for the requested predictions.**

|  |  |  |
| --- | --- | --- |
| **Company** | **Include in the request for load predictions**  **3-2.a) prediction accuracy?**  **3-2.b) a validity time?**  **3-2.c) a deadline before which predictions must be provided?**  **(More than one selection may be chosen)** | **Comment** |
| Huawei | Not sure | We understand and acknowledge the technical intention, but we are not sure if it needs to be standardized or not, since they are used for model self-evaluation. Technically, how gNB could require an accuracy of requested prediction, how the entity knows that the prediction result is accurate enough or not, if we go step further, what the entity would do if it thinks the accuracy might not be satisfied, keep working or reject, etc., then we see we make things complicated but not sure such mechanism would help or not… |
| Ericsson | 3-2.a) Yes  3-2.b) Yes  3-2.c) Needs discussions | We believe that adding accuracy information to a model output is useful to allow the receiver to understand the quality of the prediction. If the accuracy of such prediction is poor, the receiver may decide not to rely on such prediction and, e.g. to carry out load balancing based only on historical data.  The validity time is also valuable as it provides the timespan the prediction was derived for. Note that a model may be trained to provide predictions with a certain (maybe fixed) time validity. It should not be assumed that, if the Actor requests a prediction from a Model Inference function with any arbitrary reporting period, the Model Inference can generate a prediction that has a validity equal to the period.  With regards to the deadline for the requested prediction we need to understand better how this would be used. |
| Nokia | 3-2.a -> Yes  3-2.b -> No  3-2.c -> Yes | Including prediction accuracy in the request of load predictions can indicate to the node calculating those predictions how accurate those must be before it can report those back to the requesting node. In this way, also the reported information can be controlled according to the needs of the requesting node, with respect to accuracy, to use these predictions.  Regarding the validity time it is not clear to us what kind of time it is. If we are trying to predict a random variable at a certain time there are no guarantees that it will be valid in the next time instance. Validity time could be seen perhaps as expected value or as a pdf that expresses with which probability the estimate is valid as a function of time. Some explanations of the intention would be useful.  An alternative notion of time that can be defined in a deterministic manner is the time by which predictions are requested since this can be well defined and related to the needs of the requesting node, e.g., how soon it needs the predictions. This can be useful especially if it needs to coordinate load predictions from multiple nodes before making a decision. |
| Intel | For 3-2.a) and 3-2.b), we are ok to configure “whether to include validity time and prediction accuracy in the report” in the request, but not the exact value.  For 3-2.c), Yes. | In request for load prediction, it is ok to configure whether to include validity time and prediction accuracy in the report. However, for the exact threshold of report, the accuracy level and validity time may depend on the services and its QoS requirement. For example, traffic with high latency and reliability requirement may have a relatively high restrict in accuracy and validity time. Traffic with lower QoS requirement may have a relax threshold. Hence, “the accuracy the requested load predictions need to satisfy” and “a time for which the predictions must be valid” is not appropriate to be included in the request message.  However, from reporting point of view, the accuracy level and validity time should be noticed to the Actor, so that it can be aware of the how to use the received prediction result proper, e.g. whether result can be trusted, when the prediction result is valid.  We also would like to further explain that for a single message which includes prediction results from another cell may include multiple load prediction, which is predicted for a certain period. The validity time corresponding to each load prediction is used to indicate which period is this result predicted for. For example, if the neighbour gNB receives “a deadline by which predictions must be provided” (i.e. Timestamp T) in the request message, it may generate multiple load predictions, e.g. load prediction A valid from time point A to B, load prediction B valid from time point B to C, etc. Time point of each predicted result is before T. The source gNB can predict/decide based on more accurate prediction result for a certain period. |
| Vivo | 3-2.a -> not for sure  3-2.b -> Yes  3-2.c -> not for sure | For the predication with low confidence, the LB should fall back to the legacy mechanism.  The load prediction result is valid within a specific period.  We understand the intention but the function can be covered by the validity time, i.e., if the response is feedback after a long duration, then the prediction turns invalid. |
| Samsung | 3-2.a) Yes  3-2.b) Yes  3-2.c) Not sure | Since the model can not achieve 100% accuracy, whether the inference result is credible or not should be considered. The accuracy parameter may provide reference to the actors, so that the actors can adjust the decision about how to refer it accordingly, such as setting policy based on the high-accuracy inference results and taking low-accuracy results as additional reference.  The validity time is needed. The validity time is to indicate the applicative time for the results obtained from AI/ML model. For example, the predicted load should be only valid for a certain time period or time point. We need to know when the predicted value is for. Without such information, the results may not benefit to the RAN if applying it to a misplaced time.  Not sure about the “deadline”. Maybe “validity time” can be reused for this function. |
| Lenovo and Motorola Mobility | 3-2.a) Yes  3-2.b) Yes  3-2.c) Not sure | RAN node1 can indicate the demanded accuracy and validity time for load prediction to RAN node2 when requesting load prediction, RAN node 2 can feedback its load prediction when the accuracy and validity time meet the demands, e.g. RAN node 2 can provide load prediction within the validity time together with accuracy to RAN node1, thus RAN node1can make proper load balancing strategy. |
| CMCC | Need further discussion | Stage 3 details, we could discuss and agree later. We could focus on the general solution and flow chart |
| KDDI | Need further discussion | agree with CMCC. We firstly need to discuss the overall architecture and the general solution. |
| ZTE | 1. Not sure 2. OK 3. No | a)  The definition of “prediction accuracy” needs to be clearly clarified. We are not sure what “prediction accuracy” means here? Does it mean the model evaluation metrics or the accuracy of current prediction?  If it means model evaluation metrics, we think it is calculated from model training part through using testing data, but we are not sure this model evaluation metrics is useful to transfer to action with predicted information from model inference because the model evaluation metrics from cannot represent the current prediction accuracy.  If it means accuracy of current prediction, how the accuracy can be calculated before the actual labels are achieved?  b)  We are fine for this if the validity time means the certain period of prediction.  c)  Perhaps validity time includes this information if the validity time means the certain period of prediction. |
| Deutsche Telekom | 3-2a) yes  3-2b) yes  3-2c) no | 3-2a) and b) are fine with us, but their definition need further discussion.  The meaning of 3-2c) needs more clarification. |

**Summary:**

On the question whether to include in the request for load predictions

3-2.a) prediction accuracy, 3-2.b) a validity time, 3-2.c) a deadline before which predictions must be provided:

6 companies support to include 3-2.a) prediction accuracy

7 companies support to include 3-2.b) a validity time

2 companies support to include 3-2.c) a deadline before which predictions must be provided

2 companies think that we should leave these stage 3 details to later

Many companies expressed concerns what the exact definitions of accuracy, validity time and deadline are. To facilitate progress we make the following proposal.

**Proposal: It is FFS whether prediction accuracy, validity time, deadline need to be signaled in the input/output of AI/ML inference. More clarifications/definitions are needed for their definition and usability.**

## Request of Assistance Information by a gNB

In 3420 it is proposed that a gNB can request assistance information from a UE related to the user traffic, e.g., related to enhanced UE history information, data rate, packet size packet delay, next packet arrival time. Also, 4080 proposes that an NG-RAN node may use UE-assisted info, such as an MDT report with mobility history included.

**Q4.4-1 Companies are invited to provide their views on whether a gNB can request assistance information from a UE.**

|  |  |  |
| --- | --- | --- |
| **Company** | **Can a gNB request assistance information from a UE for the purpose of load prediction?** | **Comment** |
| Huawei | Not sure | On one hand, we already have MDT or L1/L2 measurement procedure to collect info from UE; on the other hand, as far as load info itself is concerned, what UE side info is needed for the gNB to evaluate the load; for mobility/trajectory, we already have UMI… |
| Ericsson | Yes | The UE has valuable information that the RAN does not have and that may be crucial to predict future load. For example, the UE might know what traffic a user may generate from the user behaviour (e.g. by knowing a facebook page is about to be loaded it is possible to predict the amount of data the UE might be exchanging in UL/DL in the next, e.g. seconds). We propose that such assistance information is explored and studied |
| Nokia | No | In our view we should not introduce new UE measurements since this is a RAN3-led SI. So, we do not support to request data rate, packet size, packet arrival time from a UE. On the other hand, since packet delay measurement already exists we could introduce it if it seems necessary. Currently, we are not sure how it can be used at the receiving NG-RAN node. On the other hand, it is unclear why we need to introduce new signalling and not re-use the MDT framework. |
| Intel | Not needed | Existing MDT procedure as baseline. For other new measurement from UE, we need to understand the solution better before defining one. |
| Vivo | Yes with comments | The existing MDT can be reused and no new UE measurement is expected. |
| Samsung | Not now | Existing MDT can be the baseline. If any new measurement is needed from detailed solution study later, it can be discussed. |
| Lenovo and Motorola Mobility |  | Agree to use the existing procedure e.g. SON/MDT for a gNB to request assistance information from a UE, but no new UE measurement. |
| CMCC | Yes | We need to identify which information is still missing from the SONMDT framework.  We could focus on the basic solution and flow chart |
| ZTE |  | We support MDT report with UE mobility history as proposed in our solution to AI-based trajectory prediction. |
| Deutsche Telekom | Not sure | We have a similar view as Samsung. Let’s first take existing MDT measurements as baseline and discuss any possible new measurements if they may provide a real benefit. |

**Summary:**

On the question can a gNB request assistance information from a UE for the purpose of load prediction. Company views are divided:

2 companies said yes,

2 companies said no,

2 companies said yes but not to introduce new UE measurements

2 companies said not to consider this topic now

1 company said yes to utilize MDT with UE mobility history

Given that this is a RAN3-led SI, moderator proposes the following:

**Proposal: If existing UE measurements are needed by a gNB for the load balancing use case, RAN3 shall consider the MDT framework as a baseline. FFS whether new RRC signaling is needed.**

# Discussion - 2nd Round

On the 2nd round of this email discussion we would like to continue the discussions on some simple FFS aspect from the 1st round and also invite comments on a draft TP on the Load Balancing use case.

Deadline for the 2nd round is Tuesday 24th, at 8am UTC.

## 5.1 Additional solution for split architecture

One more solution for split architectures is clarified from 3758 on the possible location of the AI/ML functionality, namely option 1-2.d) AI/ML training in gNB-CU and AI/ML inference in gNB-DU.

|  |  |  |
| --- | --- | --- |
| **Company** | **Do companies support the solution where AI/ML training is in gNB-CU and AI/ML inference in gNB-DU? (yes/no)** | **Comment** |
| Ericsson | Leave it for later discussions | The important part is whether the inference function can be located at the gNB-DU. Focussing on the training to be at the gNB-CU is not important as it is in a way included in the scenario where training is at the NG-RAN |
| Lenovo and Motorola Mobility | No | As answered in Q4.1-2, CU is more suitable than DU to perform AI/ML inference. |
| Samsung | Not now | We can discuss it upon the use case requirement. From current progress, it seems there is no need. |
| Intel | Yes, but | We are ok to leave it FFS and focus on the non-split architecture first. |
| ZTE | Yes | Actually, we support this solution. In short, we think gNB-CU can be responsible for the ML-training because CU may have better computation capability than DU, and then gNB-DU can be responsible for the ML inference for the real-time scenario without transferring the predicted information over F1. We are fine to discuss it later. |
| Deutsche Telekom | No | Same view as Lenovo/Motorola. Based on current use case discussion we also don’t see the need to place the inference in DU. |
| Nokia | Yes | We think that this solution can be exploited but it is ok to leave it for later. |

**Summary**

In the second phase of the discussions one more question was asked regarding the location of the AI/ML functionality in case of split architecture. In the question whether companies support a solution where AI/ML training is in gNB-CU and AI/ML inference in gNB-DU, companies answered the following:

2 companies think this should be left for later

2 companies think that this is not needed

3 companies support this solution

**It remains FFS whether we support an AI/ML Load Balancing solution where AI/ML training is in gNB-CU and AI/ML inference in gNB-DU. This is captured already in Proposal 2 above.**

## 5.2 Comments on draft TP

Companies are invited to provide comments on the draft TP for the load balancing use case.

|  |  |
| --- | --- |
| **Company** | **Comments** |
| Ericsson | The TP looks ok. The only major point of disagreement is on the use of MDT to gather Ue assistance information. This MDT based solution has been proposed by one company while there were many other companies proposing to gather UE measurements (existing or new) directly from the UE. At least we should putas FFS whether UE assistance information should be collected bv the RAN. |
| Lenovo and Motorola Mobility | In general, the TP is fine. For Figure 5.2-1, we agree that a gNB can request load predictions from a neighbouring gNB, but whether to reuse the existing procedure or introduce a new one is not agreed, we suggest to remove this figure since the messages i.e. PREDICTION STATUS REQUEST and PREDICTION STATUS RESPONSE are not agreed. |
| Samsung | The TP is fine in general. For the sentence of “Reinforcement learning”, we also prefer to remove it, as the algorithm is out of scope. And apart from it, reinforcement learning is a learning algorithm in ML and it can be used for both online and offline training instead of online only. So we prefer to remove it.  To keep consistence with framework, prefer to “Training” and “Inference” to “Model Training” and “Model Inference” respectively. |
| ZTE | Thanks for the effort to the TP and SoD. The TP is fine, but we share same view as Lenovo that although we agreed that gNBs may signal load predictions to each other over Xn interface, how to transfer the information has not been discussed yet. We can keep FFS here. |

# 6 References

|  |  |  |
| --- | --- | --- |
| [**R3-213296**](https://www.3gpp.org/ftp/TSG_RAN/WG3_Iu/TSGR3_113-e/Docs/R3-213296.zip) | Proposed TP to TR 37.817 on Load Balancing solutions and standard impacts | NEC |
| [**R3-213420**](https://www.3gpp.org/ftp/TSG_RAN/WG3_Iu/TSGR3_113-e/Docs/R3-213420.zip) | AI/ML Load Balancing use case | Ericsson |
| [**R3-213470**](https://www.3gpp.org/ftp/TSG_RAN/WG3_Iu/TSGR3_113-e/Docs/R3-213470.zip) | AI/ML based load balancing | Intel Corporation |
| [**R3-213714**](https://www.3gpp.org/ftp/TSG_RAN/WG3_Iu/TSGR3_113-e/Docs/R3-213714.zip) | Discussion on Standard Impact for RAN Intelligence (Load Balancing) | Samsung |
| [**R3-213725**](https://www.3gpp.org/ftp/TSG_RAN/WG3_Iu/TSGR3_113-e/Docs/R3-213725.zip) | Discussion on traffic load prediction | Lenovo, Motorola Mobility |
| [**R3-213758**](https://www.3gpp.org/ftp/TSG_RAN/WG3_Iu/TSGR3_113-e/Docs/R3-213758.zip) | Solution to AI based load prediction | ZTE Corporation, China Unicom |
| [**R3-213893**](https://www.3gpp.org/ftp/TSG_RAN/WG3_Iu/TSGR3_113-e/Docs/R3-213893.zip) | (TP for TR 37.817): Standards Impacts for the AI/ML Load Balancing Use Case | Nokia, Nokia Shanghai Bell |
| [**R3-214080**](https://www.3gpp.org/ftp/TSG_RAN/WG3_Iu/TSGR3_113-e/Docs/R3-214080.zip) | Further discussions on spec impacts of load balancing & traffic steering | Huawei |
| [**R3-214124**](https://www.3gpp.org/ftp/TSG_RAN/WG3_Iu/TSGR3_113-e/Docs/R3-214124.zip) | (TP to TR 37.817) Solutions for AI-based load balancing | CMCC |

# 7 Conclusion, Recommendations [if needed]

If needed