3GPP 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

[To be completed]

# 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

## 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 |
|  |  |  |

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 |
|  |  |  |

## 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. |
|  |  |  |

## 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 |
|  |  |  |

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. |
|  |  |  |

## 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 |
|  |  |  |

# 5 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 |

# 6 Conclusion, Recommendations [if needed]

If needed