**3GPP TSG-RAN WG3 Meeting #114-eR3-21xxxx**

**Online meeting, 1st Nov – 11th Nov 2021**

Agenda Item: 18.3

Source: Ericsson

Title: AI/ML Load Balancing TP

Document for: Discussion and Agreement

# TP to TR 37.817 for Load Balancing use case

# 5 Use Cases and Solutions for Artificial Intelligence in RAN

## 5.2 Load Balancing

### 5.2.1 Use case description

The rapid traffic growth and multiple frequency bands utilized in a commercial network make it challenging to steer the traffic in a balanced distribution. To address the problem, load balancing had been proposed. The objective of load balancing is to distribute load evenly among cells and among areas of cells, or to transfer part of the traffic from congested cells or from congested areas of cells, or to offload users from one cell, cell area, carrier or RAT to improve network performance. This can be done by means of optimization of handover parameters and handover actions. The automation of such optimisation can provide high quality user experience, while simultaneously improving the system capacity and also to minimize human intervention in the network management and optimization tasks.

However, the optimization of the load balancing is not an easy task as follows:

- 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.

To deal with the above issues, solutions based on AI/ML model could be introduced to improve the load balancing performance. Based on collection of various measurements and feedbacks from UEs and network nodes, historical data, etc. ML model based solutions and predicted load could improve load balancing performance, in order to provide higher quality user experience and to improve the system capacity.

* 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.

To deal with the above issues, solutions based on AI/ML model could be introduced to improve the load balancing performance. Based on collection of various measurements and feedbacks from UEs and network nodes, historical data, etc. ML model based solutions and predicted load could improve load balancing performance, in order to provide higher quality user experience and to improve the system capacity.

//////////////////////////////////////////// FIRST CHANGE /////////////////////////////////////////

### 5.2.2 Solutions and standard impacts

*Editor Note: Capture the solutions for the use case, including potential standard impacts on existing Nodes, functions, and interfaces*

The following solutions can be considered for supporting AI/ML-based load balancing:

* AI/ML Model Training is located in the OAM and AI/ML Model Inference is located in the gNB.
* AI/ML Model Training and AI/ML Model Inference are both located in the gNB.

In case of CU-DU split architecture, the following solutions are possible:

AI/ML Model Training is located in the OAM and AI/ML Model Inference is located in the gNB-CU.

AI/ML Model Training and Model Inference are both located in the gNB-CU.

The AI/ML Model Training function may consist of both online and offline training.

Other possible locations of the AI/ML Model Inference are FFS.

To improve the load balancing decisions at a gNB (gNB-CU), a gNB can request load predictions from a neighbouring node. Details of the procedure are FFS.

If existing UE measurements are needed by a gNB for AI/ML-based load balancing, RAN3 shall reuse the existing framework (including MDT and RRM measurements). FFS on whether new UE measurements are needed.

To increase the awareness of the traffic dynamics and enable more improved traffic steering decisions it is possible to complement load measurements currently exposed over RAN interfaces with information related to predicted load from neighbouring RAN nodes as well as UE measurements and information.

* An NG-RAN node can also predict its own load as well as the load of a neighbouring RAN node. This can be achieved by considering the own load and load information received from neighbour RAN nodes. RAN predictions can be signalled between RAN nodes.
* An NG-RAN node can also derive load prediction using UE measurements and information, for example MDT and RRM measurements, or UE information related to average speed per cell, and traffic. For the aspects concerning the configuration and the reporting of UE measurements and information the impacted protocol is RRC. RAN2 needs to be consulted for details during the normative phase.

Signalling of information used to derive Model Inference outputs may be achieved over the Xn interface by reusing existing or new procedures. The details are to be discussed during normative work.

An high-level signalling flow for the AI/ML use case related to Load Balancing is shown in Figure 5.2-2-1 below.



Figure 5.2-2-1: AI/ML for Load Balancing use case

Steps 1: an AI/ML Model Training is located at NG-RAN node 1. NG-RAN node 2 is assumed to have capabilities in providing NG-RAN node 1 with useful input information, such as forecasted load and/or mobility predictions.

Steps 2-3: NG-RAN node 1 can request and obtain UE measurements and information (e.g. RRM measurements, MDT measurements, average UE speed per cell, UE traffic information).

Step 4-5: NG-RAN node 1 can subscribe to receive Load information updates from the neighbouring NG-RAN node 2.

Step 6: Depending on the collected information, NG-RAN node 1 can perform Mobility Load Balancing predictions (e.g. for cells of NG-RAN node 1 and/or cells of NG-RAN node 2).

Step 7: NG-RAN node 1 takes Mobility Load Balancing decision and UEs are moved from NG-RAN node 1 to NG-RAN node 2.

Step 8: NG-RAN node 1 can subscribe to receive Feedback for Load information updates from the neighbouring NG-RAN node 2.

Step 9: NG-RAN node 2 sends Feedback for Mobility Load Balancing to NG-RAN node 1 (e.g. load updates and cell level traffic performance after load balancing).

5.2.2.1 Input of AI/ML-based Load Balancing

To predict the optimized load balancing decisions, NG-RAN may need following information as input data for AI/ML-based load balancing:

From the local node:

* Own load information (e.g. per cell, per SSB Area): these can be assumed to be some or all of the resource information in current Xn: Resource Status Update procedure
* Predicted own load information: these can be assumed to be predictions of some or all of the resource information in current Xn: Resource Status Update procedure

From the UE:

* UE location information (e.g. from RLF reports, SCG Failure Information, Successful Handover Report)
* UE Radio Measurements, e.g., RSRP, RSRQ, SINR
* UE Mobility History Information

From neighbour NG-RAN Nodes:

* Own load information (e.g. per cell, per SSB Area): these can be assumed to be some or all of the resource information in current Xn: Resource Status Update procedure
* Predicted own load information: these can be assumed to be predictions of some or all of the resource information in current Xn: Resource Status Update procedure

Editor’s Note: FFS other input information required for AI/ML-based load balancing.

5.2.2.2 Output of AI/ML-based Load Balancing

AI/ML-based load balancing model can generate following information as output:

* Selection of target cell for mobility load balancing
* Predicted own load information: these can be assumed to be predictions of some or all of the resource information in current Xn: Resource Status Update procedure
* Predicted signalled cell load information: these can be assumed to be predictions of some or all of the resource information in current Xn: Resource Status Update procedure
* Validity time for the Model Inference output predicitons

Editor’s Note: FFS other output information expected from AI/ML-based load balancing.

5.2.2.3 Feedback of AI/ML-based Load Balancing

To optimize the performance of AI/ML-based load balancing model, following feedback can be considered to be collected from NG-RAN nodes:

* UE performance information from target NG-RAN (for those UEs handed over from the source NG-RAN node)
* Load information updates from target NG-RAN
* Cell level performance after Mobility Load Balancing from target NG-RAN

Editor’s Note: FFS other feedback expected from AI/ML-based load balancing.

///////////////////////////////////////////// END OF CHANGES ///////////////////////////////////////////////

# Appendix B - TP to TR 37.817 for Mobility Optimization use case

////////////////////////////////////////////// FIRST CHANGE ////////////////////////////////////////////////

# 5 Use Cases and Solutions for Artificial Intelligence in RAN

## 5.X Mobility Optimization

### 5.X.1 Use case description

Mobility management is the scheme to guarantee the service-continuity during the mobility by minimizing the call drops, RLFs, unnecessary handovers, and ping-pong. For the future high-frequency network, as the coverage of a single node decreases, the frequency for UE to handover between nodes becomes high, especially for high-mobility UE. In addition, for the applications characterized with the stringent QoS requirements such as reliability, latency etc., the QoE is sensitive to the handover performance, so that mobility management should avoid unsuccessful handover and reduce the latency during handover procedure. However, for the conventional method, it is challengeable for trial-and-error-based scheme to achieve nearly zero-failure handover. The unsuccessful handover cases are the main reason for packet dropping or extra delay during the mobility period, which is unexpected for the packet-drop-intolerant and low-latency applications. In addition, the effectiveness of adjustment based on feedback may be weak due to randomness and inconstancy of transmission environment. Besides the baseline case of mobility, areas of optimization for mobility include dual connectivity, CHO, and DAPS, which each have additional aspects to handle in the optimization of mobility.

Mobility aspects of SON that can be enhanced by the use of AI/ML include

* Reduction of the probability of unintended events
* UE Location/Mobility/Performance prediction
* Traffic Steering

**Reduction of the probability of unintended events associated with mobility.**

Examples of such unintended events are:

* Intra-system Too Late Handover: A radio link failure (RLF) occurs after the UE has stayed for a long period of time in the cell; the UE attempts to re-establish the radio link connection in a different cell.
* Intra-system Too Early Handover: An RLF occurs shortly after a successful handover from a source cell to a target cell or a handover failure occurs during the handover procedure; the UE attempts to re-establish the radio link connection in the source cell.
* Intra-system Handover to Wrong Cell: An RLF occurs shortly after a successful handover from a source cell to a target cell or a handover failure occurs during the handover procedure; the UE attempts to re-establish the radio link connection in a cell other than the source cell and the target cell.

RAN Intelligence could observe multiple HO events with associated parameters, use this information to train its ML model and try to identify sets of parameters that lead to successful HOs and sets of parameters that lead to unintended events.

**UE Location/Mobility/Performance Prediction**

Predicting UE’s location is a key part for mobility optimisation, as many RRM actions related to mobility (e.g. selecting handover target cells) can benefit from the predicted UE location/trajectory. UE mobility prediction is also one key factor in the optimization of early data forwarding particularly for CHO. UE Performance prediction when the UE is served by certain cells is a key factor in determining which is the best mobility target for maximisation of efficiency and performance.

**Traffic Steering**

Efficient resource handling can be achieved adjusting handover trigger points and selecting optimal combination of PCell/PSCell/SCells to serve a user.

Existing traffic steering can also be improved by providing a RAN node with information related to mobility or dual connectivity.

For example, before initiating a handover, the source gNB, could use feedbacks on UE performance collected for successful handovers occurred in the past and received from neighbouring gNBs.

Similarly, for the case of dual connectivity, before triggering the addition of a secondary gNB or triggering SN change, an eNB could use information (feedbacks) received in the past from the gNB for successfully completed SN Addition or SN Change procedures.

In the two reported examples, the source RAN node of a mobility event, or the RAN node acting as Master Node (a eNB for EN-DC, a gNB for NR-DC) can use feedbacks received from the other RAN node, as input to an AI/ML function supporting traffic related decisions (e.g. selection of target cell in case of mobility, selection of a PSCell / SCell(s) in the other case), so that future decisions can be optimized.

### 5.3.2 Solutions and standard impacts

*Editor Note: Capture the solutions for the use case, including potential standard impacts on existing Nodes, functions, and interfaces*

Considering the locations of AI/ML Model Training and AI/ML Model Inference for mobility solution, following two options are considered:

* The AI/ML Model Training function is deployed in OAM, while the Model Inference function resides within the RAN node
* Both the AI/ML Model Training function and the AI/ML Model Inference function reside within the RAN node

In case of CU-DU split, Model Inference can be in the gNB-CU.

To improve the mobility decisions at a gNB (gNB-CU), a gNB can request mobility feedback from a neighbouring node. Details of the procedure are FFS.

#### 5.3.2.1 AI/ML Model Training in OAM and AI/ML Model Inference in NG-RAN node

Step 1: The RAN is assumed to have in use a trained AI/ML model for inference

Step 2. Model Inference. Required measurements are leveraged into Model Inference to output the prediction, e.g. UE trajectory prediction, target cell prediction, target NG-RAN node prediction, etc.

Step 3. According to the prediction, recommended actions or configuration are executed for Mobility Optimization.

#### 5.3.2.2 AI/ML Model Training and AI/ML Model Inference in NG-RAN node



Figure 5.3-1: Model Training and Model Inference both located in RAN node

Step 1. NG-RAN node1 trains the AI-ML mode based on the collected data and inputs received from the UEs and the neighbouring nodes. NG-RAN node 2 is assumed to have capabilities in providing NG-RAN node 1 with useful input information, such as forecasted load and/or mobility predictions.

STEP 2. NG-RAN node 1 configures the measurement information on the UE side and sends configuration message to UE requesting for radio measurement, as well as traffic and trajectory related information.UE collects the indicated measurement, e.g., UE radio related measurements such as RSRP, RSRQ, SINR of serving cell and neighbouring cells, UE traffic related information such as predicted data rate, packet size, packet delay, next packet arrival time, and the UE trajectory information including assistance information on the UE trajectory (e.g., UE speed, UE location, etc.)

Step 3. UE sends measurement report message to NG-RAN node1 including the required measurement.

Step 4. NG RAN node 1 subscribes to the Mobility Feedback Update of the neighbouring NG RAN nodes.

Step 5. NG RAN node 1 receives the Mobility Feedback Update from the neighbouring NG RAN nodes. The Mobility Feedback Update contains information including the measurements (e.g. throughput, latency, radio link quality, cell dwelling time) and the UE configurations (e.g., dual connectivity, carrier aggregation).

Step 6. NG RAN node 1 performs predictions for mobility optimization. Required measurements are leveraged into Model Inference to output the prediction, including e.g., UE trajectory prediction, target cell prediction, target NG-RAN node prediction, etc.

Step 7. According to the prediction, recommended actions are executed for Mobility Optimization e.g., handover toward NG RAN node 2.Step 8. NG RAN node 1 receives Mobility Feedback Update concerning the handed over UE from the NG RAN node 2.

#### 5.3.2.3 Input data

The following data is required as input data for mobility optimization.

**Input Information from UE:**

* FFS UE historical location information from MDT, e.g., Latitude, longitude, altitude, cell ID
* Radio measurements related to serving cell and neighbouring cells associated with UE location information, e.g., RSRP, RSRQ, SINR
* UE Mobility history information
* Moving velocity
* predicted traffic
* Trajectory information
* RAN visible QoE metrics e.g., buffer level

**Input Information from the neighbouring RAN nodes:**

* UE’s successful handover information in the past and received from neighboring RAN nodes
* UEs’ Position and trajectory, resource status, QoS parameters of historical HO-ed UE (e.g., loss rate, delay, throughput, etc.)
* Resource status and utilization prediction/estimation
* SON Reports of handovers that are successful handover report, too-early, too-late, or handover to wrong (sub-optimal) cell
* Information about the performance of handed over UEs
* UE performance prediction/estimation
* UE dwelling time per cell
* RAN visible QoE metrics e.g., buffer level

**Input Information from the local node:**

* Local load prediction

If existing UE measurements are needed by a gNB for AI/ML-based mobility optimization, RAN3 shall reuse the existing framework (including MDT and RRM measurements). FFS on whether new UE measurements are needed.

#### 5.3.2.4 Output data

* FFS UE trajectory prediction (Latitude, longitude, altitude of UE over a future period of time)
* Estimated arrival probability in CHO and relevant confidence interval
* Predicted handover target node, candidate cells in CHO, may together with the confidence of the predication

///////////////////////////////////////// END OF CHANGES ////////////////////////////////////////