3GPP TSG-RAN WG3 Meeting #113-e R3-21xxxx

E-meeting, 16– 26 August, 2021

**Agenda item: 18.4.2**

**Source: Nokia**

**Title: AI/ML Load Balancing Solution**

**Document for: Approval**

# 1 TP to TR 37.817

## 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.

-----------------------------------Start of Changes-----------------------------------

### 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 Training is located in the OAM and AI/ML inference is located in the gNB
* AI/ML Training and Inference are both located in the gNB. This can be the case if Reinforcement Learning methods are deployed in the gNB or when gNB retrains a partially trained ML model.

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

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

Other possible locations of the AI/ML functionality are FFS.

To improve the load balancing decisions at gNB (gNB-CU), a gNB can request load predictions from a neighbouring gNB. The procedure with which an NG-RAN node requests the calculation and reporting of load predictions from another NG-RAN node is shown in Figure 5.2-1.



Figure 5.2-1 Procedure with which NG-RAN nodes can request load predictions from each other. Details of the procedure are FFS.

If existing UE measurements are needed by a gNB for the load balancing use case, MDT framework shall be used as a baseline.

-----------------------------------End of Changes-----------------------------------