**3GPP TSG-RAN WG3 #115-e R3-22xxxx**

**17th Feb - 3rd Mar 2022**

**Online**

**Title:** TP to 37.817 on AI/ML based network energy saving

**Source:** ZTE

**Agenda item:** 18.4.1

**Document Type:** pCR

# 1. Introduction

This TP tries to reflect agreementS on the solution of AI/ML-based load balancing from CB # AIRAN3\_ES.

# 5. Reference

1. R3-222449, SoD\_CB # AIRAN3\_ES

# Annex – TP for TR 37.817

## 5.1 Network Energy Saving

### 5.1.1 Use case description

To meet the 5G network requirements of key performance and the demands of the unprecedented growth of the mobile subscribers, millions of base stations (BSs) are being deployed. Such rapid growth brings the issues of high energy consumption, CO2 emissions and operation expenditures (OPEX). Therefore, energy saving is an important use case which may involve different layers of the network, with mechanisms operating at different time scales.

Cell activation/deactivation is an energy saving scheme in the spatial domain that exploits traffic offloading in a layered structure to reduce the energy consumption of the whole radio access network (RAN). When the expected traffic volume is lower than a fixed threshold, the cells may be switched off, and the served Ues may be offloaded to a new target cell.

Efficient energy consumption can also be achieved by other means such as reduction of load, coverage modification, or other RAN configuration adjustments. The optimal EE decision depends on many factors including the load situation at different nodes, RAN nodes capabilities, KPI/QoS requirements, number of active Ues and UE mobility, cell utilization, etc.

However, the identification of actions aimed at energy efficiency improvements is not a trivial task. Wrong switch-off of the cells may seriously deteriorate the network performance since the remaining active cells need to serve the additional traffic. Wrong traffic offload actions may lead to a deterioration of Energy Efficiency instead of an improvement. The current energy-saving schemes are vulnerable to potential issues listed as follows:

* Inaccurate cell load prediction. Currently, energy-saving decisions rely on current traffic load without considering future traffic load.
* Conflicting targets between system performance and energy efficiency. Maximizing the system’s key performance indicator (KPI) is usually done at the expense of energy efficiency. Similarly, the most energy efficient solution may impact system performance. Thus, there is a need to balance and manage the trade-off between the two.
* Conventional energy-saving related parameters adjustment. Energy-saving related parameters configuration is set by traditional operation, e.g., based on different thresholds of cell load for cell switch on/off which is somewhat a rigid mechanism since it is difficult to set a reasonable threshold.
* Actions that may produce a local (e.g. limited to a single RAN node) improvement of Energy Efficiency, while producing an overall (e.g. involving multiple RAN nodes) deterioration of Energy Efficiency.

To deal with issues listed above, ML techniques could be utilized to leverage on the data collected in the RAN network. ML algorithms may predict the energy efficiency and load state of the next period, which can be used to make better decisions on cell activation/deactivation for ES. Based on the predicted load, the system may dynamically configure the energy-saving strategy (e.g. the switch-off timing and granularity, offloading actions) to keep a balance between system performance and energy efficiency and to reduce the energy consumption.

### 5.1.2 Solutions and standard impacts

#### 5.1.2.1 Locations for AI/ML Model Training and AI/ML Model Inference

The following solutions can be considered for supporting AI/ML-based network energy saving:

* 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.

Note: gNB is also allowed to continue model training based on AI/ML model trained in the OAM

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.

#### 5.1.2.2 Model Training at OAM and Model Inference at NG-RAN

In this solution, NG-RAN makes energy decisions using AI/ML model trained from OAM.



Figure 5.1.2.1-1. Model Training at OAM, Model Inference at NG-RAN

Step 0: NG-RAN node 2 is assumed to have an AI/ML model optionally, which can provide NG-RAN node 1 with input information.

Step 1: NG-RAN node 1 configures the measurement information on the UE side and sends configuration message to UE to perform measurement procedure and reporting.

Step 2: The UE collects the indicated measurement(s), e.g. UE measurements related to RSRP, RSRQ, SINR of serving cell and neighbouring cells.

Step 3: The UE sends the measurement report message(s) to NG-RAN node 1.

Step 4: NG-RAN node 1 further sends UE measurement reports together with other input data for Model Training to OAM.

Step 5: NG-RAN node 2 (assumed to have an AI/ML model optionally) also sends input data for Model Training to OAM. If NG-RAN node 2 executes the AI/ML model, the input data for Model Training can include the corresponding inference result from NG-RAN node 2.

Step 6: Model Training at OAM. Required measurements and input data from other NG-RAN nodes are leveraged to train AI/ML models for network energy saving.

Step 7: OAM deploys/updates AI/ML model into the NG-RAN node(s). The NG-RAN node can also continue model training based on the received AI/ML model from OAM.

Note: This step is out of RAN3 Rel-17 scope.

Step 8: NG-RAN node 2 sends the required input data to NG-RAN node 1 for model inference of AI/ML-based network energy saving.

Step 9: UE sends the UE measurement report(s) to NG-RAN node 1.

Step 10: Based on local inputs of NG-RAN node 1 and received inputs from NG-RAN node 2, NG-RAN node 1 generates model inference output(s) (e.g. energy saving strategy, handover strategy, etc).

Step 11: NG-RAN node 1 sends Model Performance Feedback to OAM if applicable.

Note: This step is out of RAN3 scope.

Step 12: NG-RAN node 1 executes Network energy saving actions according to the model inference output.

Step 13: NG-RAN node 1 may select the most appropriate target cell for each UE before it performs handover, if the output is handover strategy.

Step 14: NG-RAN node 2 provides feedback to OAM.

Step 15: NG-RAN node 1 provides feedback to OAM.

#### 5.1.2.3 Model Training and Model Inference at NG-RAN

In this solution, NG-RAN is responsible for model training and generates energy saving decisions.



Figure 5.1.2.2-1. Model Training and Model Inference at NG-RAN

Step 0: NG-RAN node 2 is assumed to have an AI/ML model optionally, which can provide NG-RAN node 1 with input information.

Step 1: NG-RAN node 1 configures the measurement information on the UE side and sends configuration message to UE to perform measurement procedure and reporting.

Step 2: The UE collects the indicated measurement(s), e.g. UE measurements related to RSRP, RSRQ, SINR of serving cell and neighbouring cells.

Step 3: The UE sends the measurement report(s) to NG-RAN node 1 including the required measurement result.

Step 4: NG-RAN node 2 sends the required input data to NG-RAN node 1 for model training of AI/ML-based network energy saving. If NG-RAN node 2 executes the AI/ML model, the input data for Model Training can include the corresponding inference result from NG-RAN node 2.

Step 5: NG-RAN node 1 trains AI/ML model for AI/ML-based energy saving based on collected data. NG-RAN node 2 is assumed to have AI/ML model for AI/ML-based energy saving optionally, which can also generate predicted results/actions.

Step 6: NG-RAN node 2 sends the required input data to NG-RAN node 1 for model inference of AI/ML-based network energy saving.

Step 7: UE sends the UE measurement report(s) to NG-RAN node 1.

Step 8: Based on local inputs of NG-RAN node 1 and received inputs from NG-RAN node 2, NG-RAN node 1 generates model inference output (e.g. energy saving strategy, handover strategy, etc).

Step 9: NG-RAN node 1 executes Network energy saving actions according to the model inference output.

Step 10: NG-RAN node 1 may select the most appropriate target cell for each UE before it performs handover, if the output is handover strategy.

Step 11: NG-RAN node 2 provides feedback to NG-RAN node 1.

#### 5.1.2.4 Input of AI/ML-based Network Energy Saving

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

Input Information from Local node:

* UE mobility/trajectory prediction
* Current/Predicted Energy efficiency
* Current/Predicted resource status

Input Information from UE:

- UE location information (e.g., coordinates, serving cell ID, moving velocity) interpreted by gNB implementation when available

* UE measurement report (e.g. UE RSRP, RSRQ, SINR measurement, etc), including cell level and beam level UE measurements

Input from neighbouring NG-RAN nodes:

* Current/Predicted energy efficiency
* Current/Predicted resource status
* Current/Predicted energy state (e.g., active, high, low, inactive)

If existing UE measurements are needed by a gNB for AI/ML-based network energy saving, RAN3 shall reuse the existing framework (including MDT and RRM measurements).

#### 5.1.2.5 Output of AI/ML-based Network Energy Saving

AI/ML-based network energy saving model can generate following information as output:

* Energy saving strategy, such as recommended cell activation/deactivation.
* Handover strategy, including recommended candidate cells for taking over the traffic
* Predicted energy efficiency
* Predicted energy state (e.g., active, high, low, inactive)
* Validity time of the predicted energy saving decisions. Validity time used outside the internal node will be discussed during R18 normative work

#### 5.1.2.6 Feedback of AI/ML-based Network Energy Saving

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

* Resource status of neighbouring NG-RAN nodes
* Energy efficiency
* UE performance affected by the energy saving action (e.g. handed-over UEs), including bitrate, packet loss, latency.
* System KPIs (e.g. throughput, delay, RLF of current and neighboring NG-RAN node)

#### 5.1.2.7 Standard Impact

MDT signalling enhancement to retrieve the AI/ML input information from UE, to be discussed during the R18 normative phase

Potential Xn interface impact:

* New signalling procedure or enhanced existing procedure to collect the input data information
	+ Predicted energy efficiency between neighboring NG-RAN nodes and source NG-RAN node
	+ Predicted resource status between neighboring NG-RAN nodes and source NG-RAN node
* New signalling procedure or enhanced existing procedure to retrieve feedback information

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