3GPP TSG-RAN WG3 Meeting #123bis R3-242160

Changsha, China, from April 15 to April 19 2024

Agenda Item: 11.2

Source: ZTE

Title: (TP to 38.743) AI/ML assisted Network Slicing

Document for: other

# 1 Introduction

This TP follows discussions in R3-24xxxx.

# 2 Text Proposal

<<<<<<<<<<<<<<<<<<<< First Change >>>>>>>>>>>>>>>>>>>>

## 4.1 AI/ML based Network Slicing

### 4.1.1 Use case description

*Editor Note: Capture the description of use case*

Network slicing is defined as a technology that logically separates network functions and resources into multiple network slices within a common physical infrastructure. Each NS represents an independent virtualized end-to-end network, providing tailored service for a specific communication scenario (e.g., eMBB, mMTC, or URLLC).

In Rel-15, 3GPP worked on network slicing advanced network architecture, providing greater flexibility and scalability for various services with different requirements. Rel-17 work focused on supporting network slicing and solutions to ensure slice-based service continuity, enabling operators to explore new sources of revenue in addition to those derived from customer subscriptions. Rel-18 work item further enhanced support of RAN slicing, including further support slice-based service continuity, solving the issue that network Slice Area of Service for services not mapping to existing TAs boundaries and temporary network slices and improved support of RAs including TAs supporting Rejected S-NSSAI.

Network slicing can provide service differentiation and ensure SLAs (Service Level Agreements) for each service type by allowing the establishment of custom slices for various types of services with different QoS requirements.

Legacy network slicing operations still fall short of ensuring the continuity of slicing services and guaranteeing QoS requirements, as outlined below:

* The NG-RAN node supports resource management between slices for the SLA assigned to each supported slice. However, relying solely on the current or historical slice-level resource status is insufficient for allocating the appropriate resources to guarantee user service quality proactively.
* For high-mobility UEs, ensuring the continuity of the sliced services is currently challenging. If the load per slice of neighboring nodes is not clearly known, the target node might not be able to provide the required service type for different users, which in turn, will cause a severe increase in outage probability. However, even if the target gNB can provide the required service type, users may experience poor achievable QoS performance due to limited resources.

The design and management of network slices require the analysis of vast amounts of complex data to effectively meet QoS requirements. Handling massive information quickly makes it challenging for manual deployment and operation of network slices. Therefore, these operations and optimizations necessitate the introduction of AI/ML techniques. AI/ML techniques can be applied to optimize the resource allocation and mobility decision related to network slicing.

AI/ML algorithms can analyze network performance, UE level and cell level traffic patterns, and UE level service requirements to determine the optimal placement of network slices and allocate resources more efficiently. By considering factors such as UE level traffic latency, bandwidth requirement, and service demands, AI/ML models can make intelligent decisions on how to allocate resources to UEs among different slices dynamically, including the scenarios of non-mobility and mobility, ensuring optimal performance and resource utilization to fulfill the requirement of SLA.

### 4.1.2 Solutions and standard impacts

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

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

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

- 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.

#### 4.1.2.2 AI/ML Model Training at OAM and AI/ML Model Inference at NG-RAN

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



Figure 4.1.2.2-1. 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.

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 slicing.

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 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 slicing.

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., network slicing 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 slicing actions according to the model inference output.

Step 13&14: NG-RAN node 1 and NG-RAN node 2 provide feedback to OAM.

#### 4.1.2.3 AI/ML Model Training and AI/ML Model Inference at NG-RAN

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



Figure 4.1.2.3-1. 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(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 slicing.

Step 5: NG-RAN node 1 trains AI/ML model for AI/ML-based network slicing based on collected data. NG-RAN node 2 is assumed to have AI/ML model for AI/ML-based network slicing 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 slicing.

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.

Step 9: NG-RAN node 1 executes network slicing actions according to the model inference output.

Step 10: NG-RAN node 2 provides feedback to NG-RAN node 1.

#### 4.1.2.4 Input data of AI/ML based Network Slicing:

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

From local node:

- Current/Predicted resource status per slice

- Current/Predicted slice available capacity

- Legacy predicted UE trajectory

From neighbouring NG-RAN nodes:

- Current/Predicted resource status per slice

- Current/Predicted slice available capacity

From the UE:

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

#### 4.1.2.5 Output data of AI/ML based Network Slicing:

AI/ML-based network slicing model in NG-RAN node can generate following information as output:

* Predicted resource status per slice
* Predicted slice available capacity
* Resource management with RRM policy (used internally)
* Slice aware mobility decisions (used internally)

#### 4.1.2.6 Feedback of AI/ML based Network Slicing:

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

* Resource Status per slice level updates from target NG-RAN node
* Slice Available Capacity updates from target NG-RAN node
* Legacy UE performance feedback for those UEs handed over from the source NG-RAN node

#### 4.1.2.7 Potential standard impacts:

Following standard impacts are listed for subsequent Rel-19 normative work compared with what was specified during Rel-18.

Xn interface:

* Enhanced existing procedure to collect predicted information between NG-RAN nodes:
  + Predicted Resource Status per slice level between neighbouring NG-RAN nodes and local NG-RAN node
  + Predicted Slice Available Capacity between neighbouring NG-RAN nodes and local NG-RAN node

<<<<<<<<<<<<<<<<<<<< End of Changes >>>>>>>>>>>>>>>>>>>>