|  |
| --- |
| 3GPP TR 37.817 V1.4.0 (2022-03) |
| Technical Report  |
| 3rd Generation Partnership Project;Technical Specification Group RAN;Evolved Universal Terrestrial Radio Access (E-UTRA) and NR;Study on enhancement for Data Collection for NR and EN-DC(Release 17) |
|   |
|  |  |
|  |
| The present document has been developed within the 3rd Generation Partnership Project (3GPP TM) and may be further elaborated for the purposes of 3GPP.The present document has not been subject to any approval process by the 3GPPOrganizational Partners and shall not be implemented.This Specification is provided for future development work within 3GPPonly. The Organizational Partners accept no liability for any use of this Specification.Specifications and Reports for implementation of the 3GPP TM system should be obtained via the 3GPP Organizational Partners’ Publications Offices. |

|  |
| --- |
|  |
| ***3GPP***Postal address3GPP support office address650 Route des Lucioles – Sophia AntipolisValbonne – FRANCETel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16Internet<http://www.3gpp.org> |
| ***Copyright Notification***No part may be reproduced except as authorized by written permission.The copyright and the foregoing restriction extend to reproduction in all media.© 2020, 3GPP Organizational Partners (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC).All rights reserved.UMTS™ is a Trade Mark of ETSI registered for the benefit of its members3GPP™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational PartnersLTE™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational PartnersGSM® and the GSM logo are registered and owned by the GSM Association |

Contents

Foreword 5

1 Scope 6

2 References 6

3 Definitions of terms, symbols and abbreviations 6

3.1 Terms 6

3.2 Symbols 6

3.3 Abbreviations 6

4 General Framework 7

4.1 High-level Principles 7

4.2 Functional Framework 7

5 Use Cases and Solutions for Artificial Intelligence in RAN 8

5.1 Network Energy Saving 8

5.1.1 Use case description 8

5.1.2 Solutions and standard impacts 9

5.1.2.1 Locations for AI/ML Model Training and AI/ML Model Inference 9

5.1.2.2 AI/ML Model Training at OAM and AI/ML Model Inference at NG-RAN 9

5.1.2.3 AI/ML Model Training and AI/ML Model Inference at NG-RAN 11

5.1.2.4 Input of AI/ML-based Network Energy Saving 12

5.1.2.5 Output of AI/ML-based Network Energy Saving 12

5.1.2.6 Feedback of AI/ML-based Network Energy Saving 13

5.1.2.7 Standard Impact 13

5.2 Load Balancing 13

5.2.1 Use case description 13

5.2.2 Solutions and standard impacts 14

5.2.2.1 Locations for AI/ML Model Training and AI/ML Model Inference 14

5.2.2.2 AI/ML Model Training in OAM and AI/ML Model Inference in a NG-RAN node 14

5.2.2.3 AI/ML Model Training and AI/ML Model Inference in a NG-RAN node 15

5.2.2.4 Input of AI/ML-based Load Balancing 17

5.2.2.5 Output of AI/ML-based Load Balancing 17

5.2.2.6 Feedback of AI/ML-based Load Balancing 17

5.2.2.7 Standard impact 17

5.3 Mobility Optimization 18

5.3.1 Use case description 18

5.3.2 Solutions and standard impacts 19

5.3.2.1 Locations for AI/ML Model Training and AI/ML Model Inference 19

5.3.2.2 AI/ML Model Training in OAM and AI/ML Model Inference in NG-RAN node 19

5.3.2.3 AI/ML Model Training and AI/ML Model Inference in a NG-RAN node 21

5.3.2.4 Input of AI/ML-based Mobility Optimization 22

5.3.2.5 Output of AI/ML-based Mobility Optimization 22

5.3.2.7 Standard impact 23

6 Conclusion 23

Annex <A> (informative): Change history 24

# Foreword

This Technical Report has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

x the first digit:

1 presented to TSG for information;

2 presented to TSG for approval;

3 or greater indicates TSG approved document under change control.

y

 the second digit is incremented for all changes of substance, i.e., technical enhancements, corrections, updates, etc.

z the third digit is incremented when editorial only changes have been incorporated in the document.

# 1 Scope

The present document provides descriptions of principles for RAN intelligence enabled by AI, the functional framework (e.g., the AI functionality and the input/output of the component for AI enabled optimization) and use cases and solutions of AI enabled RAN.

The study is based on the current architecture and interfaces.

# 2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non‑specific.

- For a specific reference, subsequent revisions do not apply.

- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] 3GPP TR 21.905: “Vocabulary for 3GPP Specifications”.

# 3 Definitions of terms, symbols and abbreviations

## 3.1 Terms

For the purposes of the present document, the terms given in 3GPP TR 21.905 [1] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in 3GPP TR 21.905 [1].

* Data collection: Data collected from the network nodes, management entity or UE, as a basis for AI/ML model training, data analytics and inference.
* AI/ML Model: A data driven algorithm by applying machine learning techniques that generates a set of outputs consisting of predicted information and/or decision parameters, based on a set of inputs
* AI/ML Training: An online or offline process to train an AI/ML model by learning features and patterns that best present data and get the trained AI/ML model for inference.
* AI/ML Inference: A process of using a trained AI/ML model to make a prediction or guide the decision based on collected data and AI/ML model.

## 3.2 Symbols

For the purposes of the present document, the following symbols apply:

<symbol> <Explanation>

## 3.3 Abbreviations

For the purposes of the present document, the abbreviations given in 3GPP TR 21.905 [1] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in 3GPP TR 21.905 [1].

<ABBREVIATION> <Expansion>

# 4 General Framework

## 4.1 High-level Principles

The following high-level principles should be applied for AI-enabled RAN intelligence:

* The detailed AI/ML algorithms and models for use cases are implementation specific and out of RAN3 scope.
* The study focuses on AI/ML functionality and corresponding types of inputs/outputs.
* The input/output and the location of the Model Training and Model Inference function should be studied case by case.
* The study focuses on the analysis of data needed at the Model Training function from Data Collection, while the aspects of how the Model Training function uses inputs to train a model are out of RAN3 scope.
* The study focuses on the analysis of data needed at the Model Inference function from Data Collection, while the aspects of how the Model Inference function uses inputs to derive outputs are out of RAN3 scope.
* Where AI/ML functionality resides within the current RAN architecture, depends on deployment and on the specific use cases.
* The Model Training and Model Inference functions should be able to request, if needed, specific information to be used to train or execute the AI/ML algorithm and to avoid reception of unnecessary information. The nature of such information depends on the use case and on the AI/ML algorithm.
* The Model Inference function should signal the outputs of the model only to nodes that have explicitly requested them (e.g., via subscription), or nodes that take actions based on the output from Model Inference.
* An AI/ML model used in a Model Inference function has to be initially trained, validated and tested by the Model Training function before deployment.
* NG-RAN SA is prioritized; EN-DC and MR-DC are down-prioritized, but not precluded from Rel.18.
* Functional framework and high-level procedures defined in this TR should not prevent from “thinking beyond” them during normative phase if a use case requires so.
* User data privacy and anonymisation should be respected during AI/ML operation.

## 4.2 Functional Framework



Figure 4.2-1. Functional Framework for RAN Intelligence

This section introduces the common terminologies related to the functional framework for RAN intelligence illustrated in Figure 4.2-1. For the functions and data/information flows shown in the Figure 4.2-1, whether there is any standardization impact and what is the standardization impact are discussed in clause 5.

* Data Collection is a function that provides input data to Model training and Model inference functions. AI/ML algorithm specific data preparation (e.g., data pre-processing and cleaning, formatting, and transformation) is not carried out in the Data Collection function.
Examples of input data may include measurements from UEs or different network entities, feedback from Actor, output from an AI/ML model.
	+ Training Data: Data needed as input for the AI/ML Model Training function.
	+ Inference Data: Data needed as input for the AI/ML Model Inference function.
* Model Training is a function that performs the AI/ML model training, validation, and testing which may generate model performance metrics as part of the model testing procedure. The Model Training function is also responsible for data preparation (e.g., data pre-processing and cleaning, formatting, and transformation) based on Training Data delivered by a Data Collection function, if required.
	+ Model Deployment/Update: Used to initially deploy a trained, validated, and tested AI/ML model to the Model Inference function or to deliver an updated model to the Model Inference function.
		- Note: Details of the Model Deployment/Update process as well as the use case specific AI/ML models transferred via this process are out of RAN3 Rel-17 study scope. The feasibility to single-vendor or multi-vendor environment has not been studied in RAN3 Rel-17 study.
* Model Inference is a function that provides AI/ML model inference output (e.g., predictions or decisions). Model Inference function may provide Model Performance Feedback to Model Training function when applicable. The Model Inference function is also responsible for data preparation (e.g., data pre-processing and cleaning, formatting, and transformation) based on Inference Data delivered by a Data Collection function, if required.
	+ Output: The inference output of the AI/ML model produced by a Model Inference function.
		- Note: Details of inference output are use case specific.
	+ Model Performance Feedback: It may be used for monitoring the performance of the AI/ML model, when available.
		- Note: Details of the Model Performance Feedback process are out of RAN3 scope.
* Actor is a function that receives the output from the Model Inference function and triggers or performs corresponding actions. The Actor may trigger actions directed to other entities or to itself.
	+ Feedback: Information that may be needed to derive training data, inference data or to monitor the performance of the AI/ML Model and its impact to the network through updating of KPIs and performance counters.

# 5 Use Cases and Solutions for Artificial Intelligence in RAN

## 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 energy saving decision depends on many factors including the load situation at different RAN 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 optimize the energy saving decisions by leveraging 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 energy saving. 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 AI/ML Model Training at OAM and AI/ML 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.

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. 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 13: NG-RAN node 2 provides feedback to OAM.

Step 14: NG-RAN node 1 provides feedback to OAM.

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

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

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. 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 10: 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:

From local node:

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

From the 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

From neighbouring NG-RAN nodes:

* Current/Predicted energy efficiency
* Current/Predicted resource status
* Current 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)
* Model output validity time will be discussed during R18 normative work per inference output.

#### 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 neighbouring NG-RAN node)

#### 5.1.2.7 Standard Impact

MDT procedure enhancements should be discussed during the normative phase.

Potential Xn interface impact:

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

## 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. AI/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.

### 5.2.2 Solutions and standard impacts

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

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.

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

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

A high-level signalling flow for the AI/ML use case related to Load Balancing with Model Training in OAM and Model Inference in NG-RAN is shown in Figure 5.2.2-1 below.



 Figure 5.2.2-1. 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 useful input information, such as predicted resource status, etc.

Step 1: The NG-RAN node 1 configures the UE to provide measurements and/or location information (e.g., RRM measurements, MDT measurements, velocity, position).

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 reports to NG-RAN node 1 requested measurements and/or location information (e.g., UE measurements related to RSRP, RSRQ, SINR of serving cell and neighbouring cells, velocity, position).

Step 4: NG-RAN node 1 further sends UE measurement reports together with other input data for Model Training to OAM. NG-RAN node 2 also sends input data for Model Training to OAM.

Step 5: AI/ML Model Training is located at OAM. The required measurements and input data from other NG-RAN nodes are leveraged to train the AI/ML model.

Step 6: OAM deploys/updates AI/ML model into the NG-RAN node(s). The NG-RAN node is allowed to continue model training based on the received AI/ML model from OAM.

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

Step 7: The UE collects and reports to NG-RAN node 1 requested measurements or location information.

Step 8: The NG-RAN node 1 receives from the neighbouring NG-RAN node 2 the input information for load balancing model inference.

Step 9: NG-RAN node 1 performs model inference and generate Load Balancing predictions or decisions.

Step 10. The NG-RAN 1 sends the model performance feedback to OAM if applicable.

Note: This step is out of RAN3 scope.

Step 11: NG-RAN node 1 may take Load Balancing actions and the UE is moved from NG-RAN node 1 to NG-RAN node 2.

Step12: NG-RAN node 1 and NG-RAN node 2 send feedback information to OAM.

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

A high-level signalling flow for the AI/ML use case related to Load Balancing with Model Training and Model Inference in a NG-RAN node is shown in Figure 5.2.2-2 below.



Figure 5.2.2-2. Model Training and Model Inference in a NG-RAN node

Step 0: NG-RAN node 2 is assumed to have an AI/ML model optionally, which can provide NG-RAN node 1 with useful input information, such as predicted resource status, etc.

Step 1: The NG-RAN node 1 configures UE to provide measurements and/or location information(e.g., RRM measurements, MDT measurements, velocity, position).

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

Step 3: The UE reports to NG-RAN node 1 the requested measurements and/or location information (e.g., UE measurements related to RSRP, RSRQ, SINR of serving cell and neighbouring cells, velocity, position).

Step 4: The NG-RAN node 1 receives from the neighbouring NG-RAN node 2 the input information for load balancing model training.

Step 5: An AI/ML Model Training is located at NG-RAN node 1. The required measurements and input data from other NG-RAN nodes are leveraged to train the AI/ML model.

Step6: NG-RAN node 1 receives UE measurements and/or location information.

Step7: NG-RAN node 1 can receive from the neighbouring NG-RAN node 2 the input information for load balancing model inference.

Step 8: NG-RAN node 1 performs model inference and generate Load Balancing predictions or decisions.

Step 9: NG-RAN node 1 may take Load Balancing actions and the UE is moved from NG-RAN node 1 to NG-RAN node 2.

Step 10: NG-RAN node 2 sends feedback information to NG-RAN node 1 (e.g., resource status updates after load balancing, etc).

#### 5.2.2.4 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:

* Current and predicted own resource status
* UE trajectory prediction
* Current and predicted UE traffic
* Predicted resource status information of neighbouring NG-RAN node(s)

From the UE:

* UE location information (e.g., coordinates, serving cell ID, moving velocity) interpreted by gNB implementation when available
* UE Mobility History Information
* UE measurement report (e.g., UE RSRP, RSRQ, SINR measurement, etc), including cell level and beam level UE measurements

From neighbouring NG-RAN Nodes:

* Current and predicted resource status
* UE performance measurement at traffic offloaded neighbouring cell

#### 5.2.2.5 Output of AI/ML-based Load Balancing

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

* Selection of target cell for load balancing
* Predicted own resource status information
* Predicted resource status information of neighbouring NG-RAN node(s)
* Model output validity time will be discussed during R18 normative work per inference output.
* The predicted UE(s) selected to be handed over to target NG-RAN node (will be used by RAN node internally)

#### 5.2.2.6 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)
* Resource status information updates from target NG-RAN
* System KPIs (e.g., throughput, delay, RLF of current and neighbours)

#### 5.2.2.7 Standard impact

To improve the load balancing decisions at a gNB, a gNB can request load predictions from a neighbouring node. Details of the procedure will be determined during the normative phase.

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). Whether new UE measurements are needed is left to normative phase based on the use case description.

MDT procedure enhancements should be discussed during the normative phase.

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. This can be achieved by considering the own load and load information received from neighbour RAN nodes. Load 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 location information (e.g., velocity, position). 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.

**Potential Xn interface impact:**

* New or enhanced existing signaling procedure to request/retrieve predicted resource status information from neighbouring nodes via Xn interface.
* New or enhanced existing signaling procedure to request/retrieve predicted load balancing strategy information from neighbouring nodes via Xn interface.
* New or enhanced existing procedure to request/retrieve feedback information via Xn interface.

## 5.3 Mobility Optimization

### 5.3.1 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 has 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.
* Successful Handover: During a successful handover, there is underlying issue.

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

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

Considering the locations of AI/ML Model Training and AI/ML Model Inference for mobility solution, the 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

Furthermore, for CU-DU split scenario, following option is possible:

* AI/ML Model Training is located in CU-CP or OAM, and AI/ML Model Inference function is located in CU-CP

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

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



 Figure 5.3-1. AI/ML Model Training in OAM and AI/ML Model Inference in NG-RAN node

Step 0. NG-RAN node 2 is assumed to optionally have an AI/ML model, which can generate required input such as resource status and utilization prediction/estimation etc.

Step 1. The NG-RAN node configures the measurement information on the UE side and sends configuration message to UE including configuration information.

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

Step 3. The UE sends measurement report message to NG-RAN node 1 including the required measurement.

Step 4. The NG-RAN node 1 sends the input data for training to OAM, where the input data for training includes the required input information from the NG-RAN node 1 and the measurement from UE.

Step 5. The NG-RAN node 2 sends the input data for training to OAM, where the input data for training includes the required input information from the NG-RAN node 2. If the NG-RAN node 2 executes the AI/ML model, the input data for training can include the corresponding inference result from the NG-RAN node 2.

Step 6. Model Training. Required measurements are leveraged to training AI/ML model for UE mobility optimization.

Step 7. OAM sends AI/ML Model Deployment Message to deploy the trained/updated 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. The NG-RAN node 1 obtains the measurement report as inference data for UE mobility optimization.

Step 9. The NG-RAN node 1 obtains the input data for inference from the NG-RAN node 2 for UE mobility optimization, where the input data for inference includes the required input information from the NG-RAN node 2. If the NG-RAN node 2 executes the AI/ML model, the input data for inference can include the corresponding inference result from the NG-RAN node 2.

Step 10. 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 11. The NG-RAN 1 sends the model performance feedback to OAM if applicable.

Note: This step is out of RAN3 scope.

Step 12: According to the prediction, recommended actions or configuration, the NG-RAN node 1, the target NG-RAN node (represented by NG-RAN node 2 of this step in the flowchart), and UE perform the Mobility Optimization / handover procedure to hand over UE from NG-RAN node 1 to the target NG-RAN node.

Step 13. The NG-RAN node 1 sends the feedback information to OAM.

Step 14. The NG-RAN node 2 sends the feedback information to OAM.

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



Figure 5.3-2. Model Training and Model Inference both located in RAN node

Step 0. NG-RAN node 2 is assumed to optionally have an AI/ML model, which can generate required input such as resource status and utilization prediction/estimation etc.

Step 1. NG-RAN node1 configures the measurement information on the UE side and sends configuration message to UE including configuration information.

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

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

Step 4. The NG-RAN node 1 obtains the input data for training from the NG-RAN node2, where the input data for training includes the required input information from the NG-RAN node 2. If the NG-RAN node 2 executes the AI/ML model, the input data for training can include the corresponding inference result from the NG-RAN node 2.

Step 5. Model training. Required measurements are leveraged to training AI/ML model for mobility optimization.

Step 6. NG-RAN node1 obtains the measurement report as inference data for real-time UE mobility optimization.

Step 7. The NG-RAN node 1 obtains the input data for inference from the NG-RAN node 2 for UE mobility optimization, where the input data for inference includes the required input information from the NG-RAN node 2. If the NG-RAN node 2 executes the AI/ML model, the input data for inference can include the corresponding inference result from the NG-RAN node 2.

Step 8. Model Inference. 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 9: According to the prediction, recommended actions or configuration, the NG-RAN node 1, the target NG-RAN node (represented by NG-RAN node 2 of this step in the flowchart), and UE perform the Mobility Optimization / handover procedure to hand over UE from NG-RAN node 1 to the target NG-RAN node.

Step 10. The NG-RAN node 2 sends feedback information after mobility optimization action to the NG-RAN node 1.

Note: UE mobility information for training purposes is only sent to gNBs that requested such information or when triggered.

#### 5.3.2.4 Input of AI/ML-based Mobility Optimization

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

From the UE:

* UE location information (e.g., coordinates, serving cell ID, moving velocity) interpreted by gNB implementation when available.
* Radio measurements related to serving cell and neighbouring cells associated with UE location information, e.g., RSRP, RSRQ, SINR.
* UE Mobility History Information.

From the neighbouring RAN nodes:

* UE’s history information from neighbour
* Position, QoS parameters and the performance information of historical HO-ed UE (e.g., loss rate, delay, etc.)
* Current/predicted resource status
* UE handovers in the past that were successful and unsuccessful, including too-early, too-late, or handover to wrong (sub-optimal) cell, based on existing SON/RLF report mechanism.

From the local node:

* UE trajectory prediction
* Current/predicted resource status
* Current/predicted UE traffic

#### 5.3.2.5 Output of AI/ML-based Mobility Optimization

AI/ML-based mobility optimization can generate following information as output:

* UE trajectory prediction (Latitude, longitude, altitude, cell ID of UE over a future period of time)

Note: Whether the UE trajectory prediction is an external output to the node hosting the Model Inference function should be discussed during the normative work phase.

* 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
* Priority, handover execution timing, predicted resource reservation time window for CHO.
* UE traffic prediction (will be used by the RAN node internally and the details are left to normative work phase)
* Model output validity time will be discussed during R18 normative work per inference output.

5.3.2.6 Feedback of AI/ML-based Mobility Optimization

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

* QoS parameters such as throughput, packet delay of the handed-over UE, etc.
* Resource status information updates from target NG-RAN.
* Performance information from target NG-RAN. The details of performance information are to be discussed during normative work phase.

#### 5.3.2.7 Standard impact

To improve the mobility decisions at a gNB (gNB-CU), a gNB can request mobility feedback from a neighbouring node. Details of the procedure will be determined during the normative phase.

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). Whether new UE measurements are needed is left to normative phase based on the use case description.

MDT procedure enhancements should be discussed during the normative phase.

* **Potential Xn interface impact:**
	+ Predicted resource status info and performance info from candidate target NG-RAN node to source NG-RAN node
	+ New signaling procedure or existing procedure to retrieve input information via Xn interface.
	+ New signaling procedure or existing procedure to retrieve feedback information via Xn interface.

# 6 Conclusion

The AI/ML functionality and the following use cases are recommended by RAN3 to be specified in Rel-18 normative phase:

* AI/ML-based Network Energy Saving
* AI/ML-based Load Balancing
* AI/ML-based Mobility Optimization

Recommendations for each use case take the section of “Solutions and standard impacts” for each use case as basis. The high-level principles captured in section 4.1 of TR37.817 shall remain valid during normative phase, while the functional framework captured in section 4.2 of TR37.817 should be used as a guideline in normative phase.

Annex <A> (informative):
Change history

|  |
| --- |
| **Change history** |
| **Date** | **Meeting** | **Tdoc** | **CR** | **Rev** | **Cat** | **Subject/Comment** | **New version** |
| 2020-11 | RAN3#110 | R3-207094 | - | - | - | Draft skeleton | 0.0.0 |
| 2020-11 | RAN3#110 | R3-207253 |  |  |  | Capture TP in R3-207218 | 0.1.0 |
| 2021-05 | RAN3#112 | R3-212990 |  |  |  | Capture TP in R3-212807, R3-212868, R3-212896, R3-212897, R3-212978 | 0.2.0 |
| 2021-08 | RAN3#113 | R3-214517 |  |  |  | Capture TP in R3-214481, R3-214482, R3-214483, R3-214484 | 0.3.0 |
| 2021-11 | RAN3#114 | R3-216278 |  |  |  | Capture TP in R3-216192, R3-216228, R3-216230, R3-216232 | 0.4.0 |
| 2021-12 | RAN#94 | RP-213048 |  |  |  | Submit to RAN#94 | 1.0.0 |
| 2022-01 | RAN3#114b | R3-221014 |  |  |  | Resubmission of v1.0.0 | 1.1.0 |
| 2022-01 | RAN3#114b | R3-221610 |  |  |  | Capture TP in R3-221221, R3-221440, R3-221446, R3-221467 | 1.2.0 |
| 2022-03 | RAN3#115 | R3-222969 |  |  |  | Capture TP in R3-222764, R3-222798, R3-222800, R3-222865, R3-222866 | 1.3.0 |
| 2022-03 | RAN3#115 | R3-222989 |  |  |  | Update the figure 5.3-2 | 1.4.0 |