AI/ML for RAN

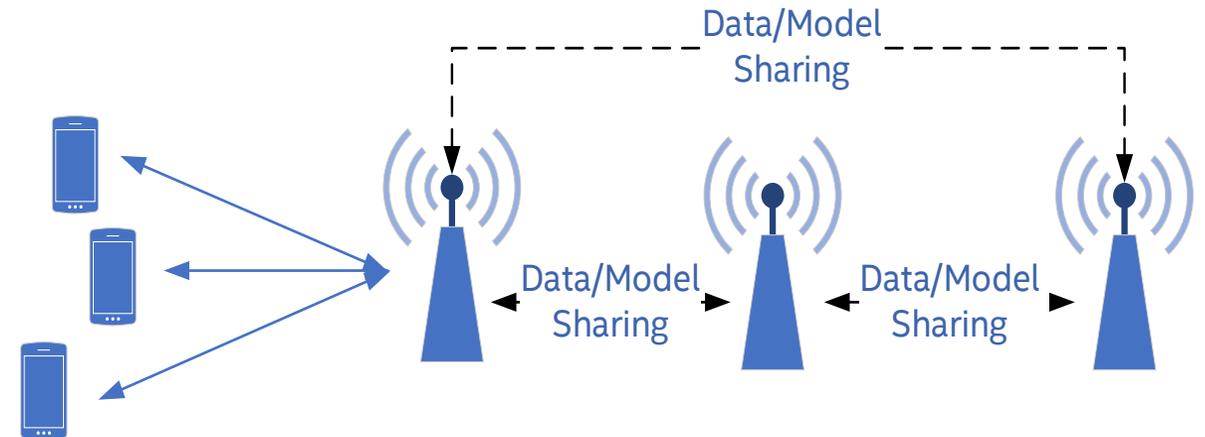# Further enhancement for data collection

- Continue work item based on outcome from SI: further enhancement for data collection (RP-201304)

- Current status of SI:

  - Support generic functional framework for NG-RAN AI/ML, supporting AI/ML automation and data collection on top of SON/MDT.

  - AI/ML functions (e.g., ML training/ML inference) are located at network side, such as RAN, OAM, CN.

  - Interface and standards impacts mainly focus on data collection and exchange over F1, NG, Xn interfaces.

  - Use cases include energy saving, load balancing, mobility management.

- Focusing on standardization of use cases: energy saving, load balancing, traffic steering/mobility optimization. [RAN3]

  - To standardize data collection as input/output for "Model Training" and "Model Inference".

  - To standardize interface to support information exchanging among network nodes and AI functions.

Supporting wider range of use cases in NG-RAN

▪ Only SON/MDT related use cases (e.g., energy saving, load balancing, traffic steering/mobility optimization) were prioritized in Rel-17 SI. Other network-level use cases can also benefit from AI/ML:

- Network slicing: specify the mechanism to optimize network resource allocation, cell selection, service continuity, etc.

- RAN QoE optimization: specify the mechanism to optimize network resource allocation by using QoE measurement result, RRM measurement, etc.

- IAB: specify the mechanism to optimize resource allocation, topology adaptation of each IAB-node in the network.

▪ Study the benefits of AI-enabled use cases, e.g., network slicing, RAN QoE optimization, IAB network, etc [RAN2, RAN3]:

- Study and evaluate the standardization impacts for the identified use cases, including data collection, input/output from Model Training and Model Inference

- Study and specify the node or function to receive/provide the input/output data

- Study and specify the network interface(s) to convey the required data among AI functions.

Distributed/Federated AI-enabled NG-RAN

- Instead of training a well-designed AI/ML model via massive input data, accuracy of AI/ML model can also be improved by utilizing other network nodes' AI/ML models.

- Enhancement to improve AI/ML model accuracy including [RAN3]

  - Study and evaluate the model performance under distributed and federated AI-enabled NG-RAN.

  - Study and evaluate the standardization impacts for data and model sharing between NG-RAN.
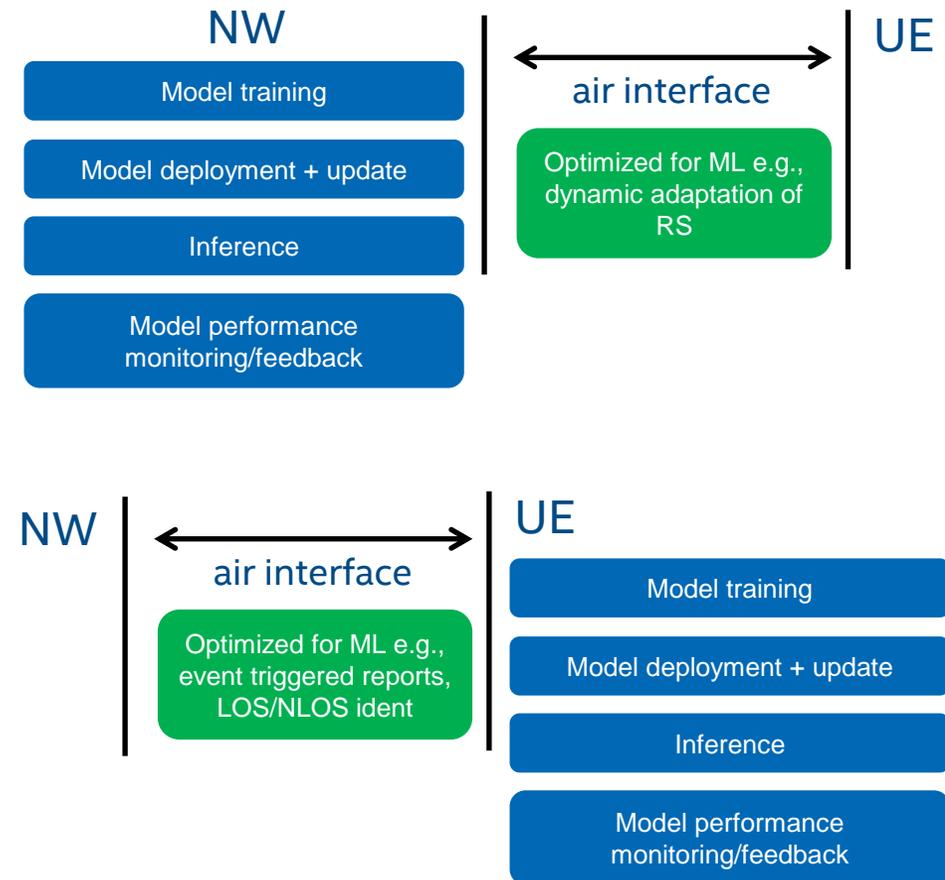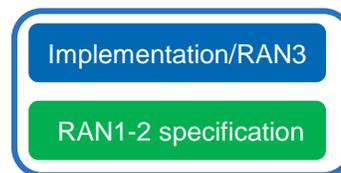
AI/ML for Air Interface

AI/ML for Air Interface

- Considering AI/ML functional framework agreed in TR 37.817, network and UE can act as different function node within the framework.

Step 1

- Model Training, Model Inference and Data Collection locate at gNB, UE can act as the Actor which is responsible for perform action according to network configuration (i.e. output from ML Inference).
- The UE can provide feedback (e.g., performance feedback, measurement report, etc.) to the network as close-loop for ML model optimization.

Step 2

- The UE also have Model Training, Model Inference function node. UE can request data for its own model's training or inference from the network. The requested data can be original data or output from NG-RAN's own training/inference output.
- The UE decides its own action according to the output from UE's Model Inference function node. It can also report the feedback or action to the network, if needed.

Air interface impact is limited to support assistance information for ML training and ML inference at a single end

Step 3

- Network acts as the Model Training function node, and the ML inference model is deployed to the UE over the air interface. The UE performs ML inference and act according to inference outcome.
- ML model can be deployed or updated according to the feedback (e.g., model performance feedback, performance feedback, etc.) provided from UE to the network.

Step 4

- Both network and the UE have the capability as Model training and Model inference function node, where ML model at the UE is downloaded/updated from the network. The network acts as the centralized node for the general ML model training.
- The UE updates/re-fines ML model according to its own environment and performance. The UE also provides ML model update information (e.g., model parameters, etc) back to the network for the general ML model update.

Air interface impact is significant to support model deployment/update and model parameter update.
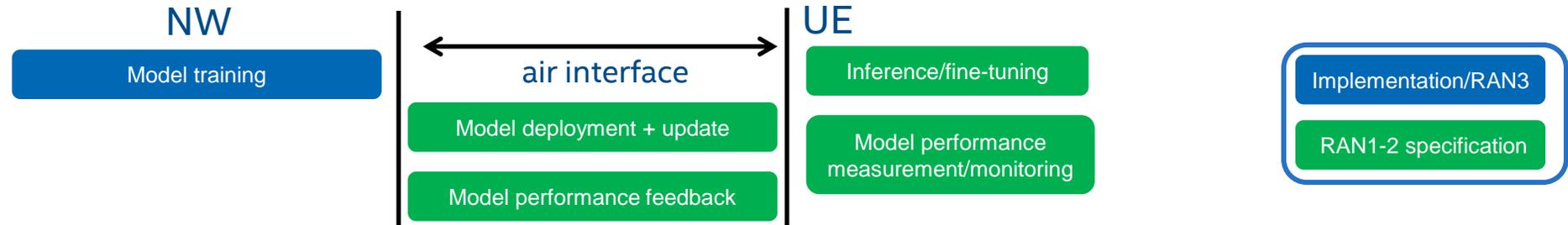
Initial enhancements

- Model training and inference <u>is not</u> spread across the air-interface (steps 1-2)

  - No need to specify ML model/parameters or performance feedback loop

  - No need to specify ML related device capabilities

  - ML is implemented according to the device's individual hardware capabilities (AI accelerators)

NW ───────────── **UE**

← air interface →

NW
Model training
Model deployment + update
Inference
Model performance monitoring/feedback

UE
Optimized for ML e.g., dynamic adaptation of RS

NW ───────────── **UE**

← air interface →

NW
Optimized for ML e.g., event triggered reports, LOS/NLOS ident

UE
Model training
Model deployment + update
Inference
Model performance monitoring/feedback

Implementation/RAN3
RAN1-2 specification

Advanced enhancements

- ▪ Model training and inference <u>is</u> spread across the air-interface (steps 3-4)

  - • Scope of evaluations and specifications is significantly higher

  - • ML hardware capabilities (accelerators) vastly differ in smartphones

  - • Inferencing at slot/sub-slot timing level (or very large models) involves model optimization for specific UE hardware (quantization, compression)

  - • Consideration of UE implementation impairments, ML model complexity (based on UE capability) becomes more critical

Example use-cases – MIMO

- Initial enhancements

  - NN at NW

    - Dynamic DM-RS/CSI-RS/SRS pattern adaptation (e.g., number of additional DM-RS), TCI switching prediction, CSI acquisition by gNB using low-overhead CSI-RS/feedback
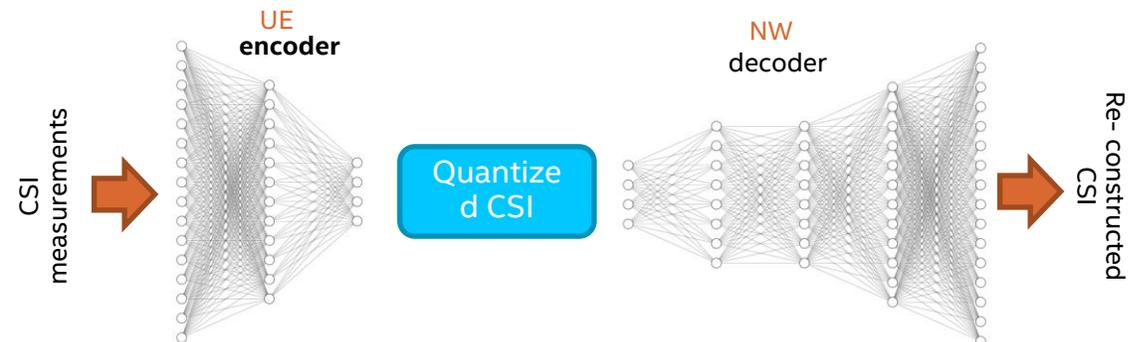
  - NN at UE

    - Low latency beam acquisition, UE feedback for adaptation of Tx parameters (e.g., DM-RS parameters, interference prediction), event-based CSI reporting

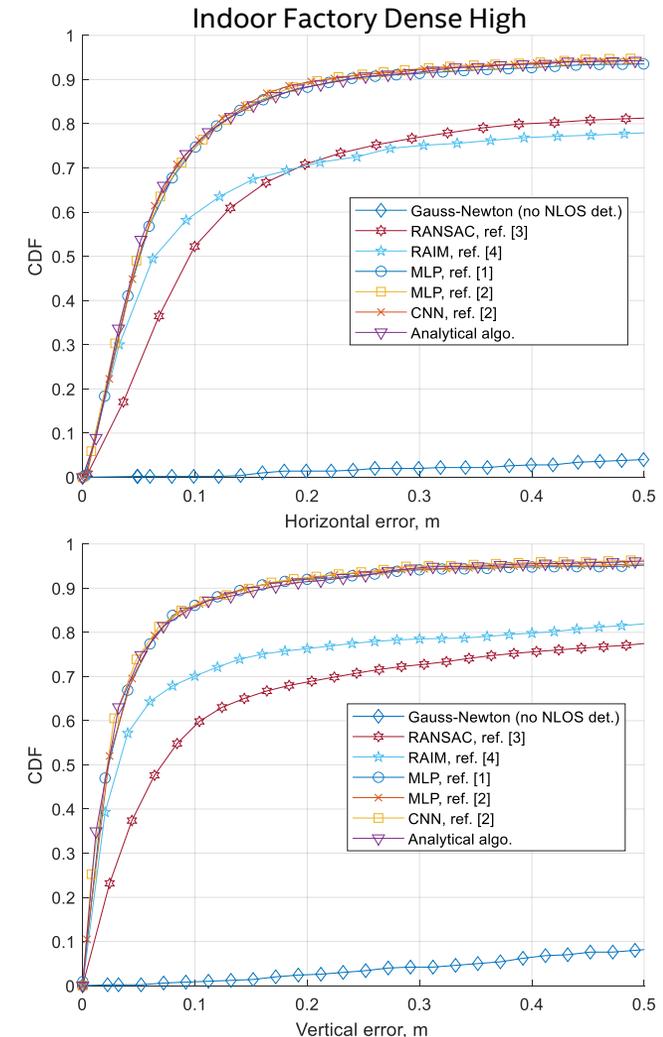- Advanced enhancements

  - NN at NW + UE

    - CSI compression to assist DL precoding, enhanced TCI switching w/o signalling assistance

CSI measurements → UE encoder → Quantized CSI → NW decoder → Re-constructed CSI

Example use-cases - Positioning

- Initial enhancements

  - NN at UE

    - UE reporting of LOS/NLOS link identification for mitigation of accuracy loss due to multipath/NLOS links.

    - Multi Layer Perceptron (MLP), [1]):

    - MLP, (ref. [2])

    - CNN, (ref. [2])

- ML methods allows to achieve the positioning accuracy of 0.2 m in both horizontal and vertical dimensions for 90 % of users

Note: Ref [1],[2],[3],[4] in results refer to references [4],[5],[2],[3] respectively in R1-2104909

Indoor Factory Dense High