3GPP TSG RAN Meeting #93e

RP-212340

Electronic Meeting, 13-17 September 2021

# AI/ML enabled RAN and NR Air Interface

Agenda Item: 9.0.2

Source: Intel Corporation

Document for: Discussion

intel®

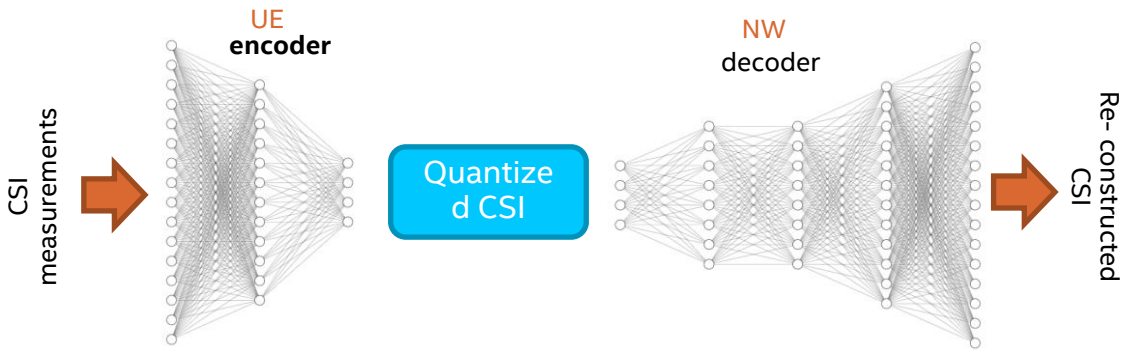
# AI/ML for Air Interface

# High level views

- Focused scope
  - Important to define a focused air-interface SI scope, otherwise WG discussions can be inefficient.
- Use case down-selection
  - High-level use-case selection in RAN plenary is preferred for air-interface (it is complex and time-consuming to do a similar exercise in RAN1 as use-cases are coupled with EVM and specifications)
- Basic ML evaluations
  - focus on offline and centralized learning evaluations
- Early RAN4 involvement
  - Consider methodologies for performance requirements and testing [RAN4] for AI/ML. Some discussion should start in RAN4 within the air-interface SI term.

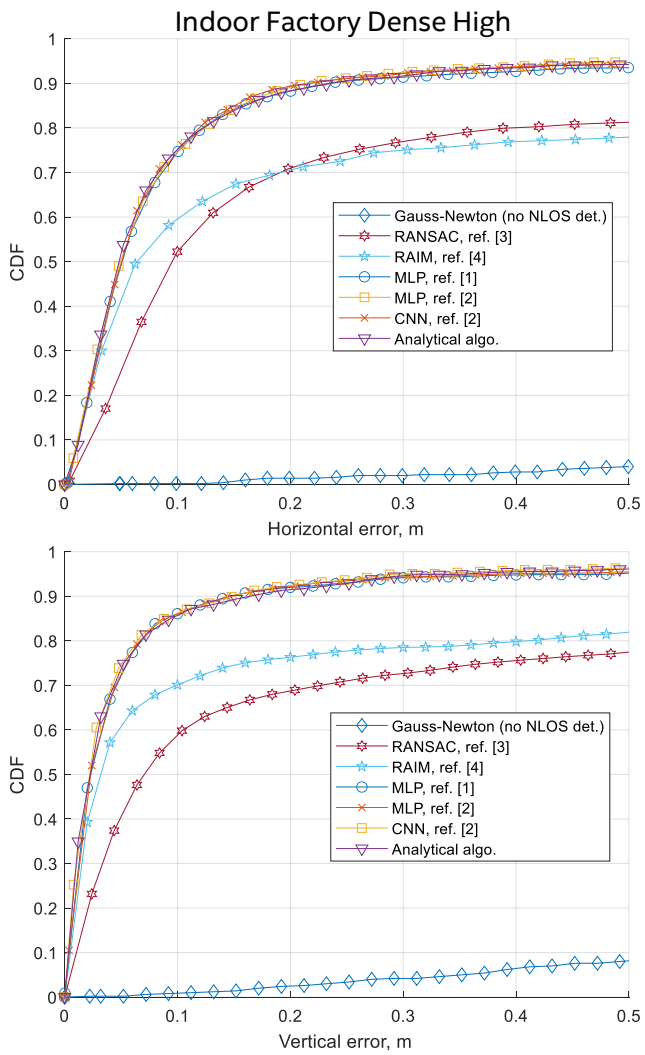
# Use cases - MIMO

- Beam management
  - Fast beam-pair acquisition on downlink for single TRP operation
- RS overhead reduction
  - Overhead reduction for CSI-RS for CSI
- CSI feedback
  - Un-precoded CSI-RS based rank and precoder feedback. Focus on large Tx ports. Spatial, frequency and time-domain compression can be considered. Rel-16 Type-II as a baseline.



# Use cases - Positioning

- LOS/NLOS classification and multi-path mitigation for DL positioning
- Study assuming current architecture (measurements at UE/gNB side, reported to LMF)
- Note: Ref [1],[2],[3],[4] in results refer to references [4],[5],[2],[3] respectively in R1-2104909



# Inter-node collaboration

- Model training and inference is not spread across the air-interface
  - 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)

- Model training and inference is spread across the air-interface
  - 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

# KPIs and EVM

- RAN1 should define methodology in order to evaluate robustness of ML solutions (e.g., minimum performance guarantee)
- Some positioning problems for coordinate inference may require channel model enhancements for spatial consistency (statistical models are not suitable)
    - we think this may not be feasible within the work scope and this issue should be considered when defining use-cases

# AI/ML for RAN

# High level views

- Continue on WI after RAN3 Rel-17 SI
  - Normative work based on the outcome of RAN3-led Rel-17 SI *FS_NR_ENDC_data_collect*.
- New RAN3-led Rel-18 SI
  - Study additional use cases of AI/ML for NG-RAN, e.g. QoE, URLLC, network slicing, IAB, etc.
  - Study further architecture and framework enhancement

# Further enhancement for data collection

- Current status of Rel-17 Study Item
  - Generic functional framework for NG-RAN AI/ML, supporting network automation and data collection on top of SON/MDT.
  - Identifying required input/output of AI/ML model and their standard's impacts.
  - Use cases include energy saving, load balancing, mobility management.
  - AI/ML functions (e.g., ML training/ML inference) are located at network side, such as RAN, OAM.

- Rel-18 Work Item based on the outcome from the Rel-17 SI
  - Normative works for energy saving, load balancing, mobility optimization use cases [RAN3]
    - Standardize required input data for model training and interference
    - Standardize expected output data or behavior from model training and interference
    - Standardize information exchange between network nodes and AI functions

# New Rel-18 SI: more use cases in NG-RAN

- Only SON/MDT related use cases (energy saving, load balancing, mobility optimization) were prioritized in Rel-17 SI. Other network-level use cases can also benefit from AI/ML:
  - Network Slicing: network resource allocation, cell selection, service continuity, etc.
  - RAN QoE: network resource allocation by using QoE measurement result, RRM measurement, etc.
  - IAB: resource allocations across IAB nodes, topology adaptation of each IAB-node in the network
- Study additional use cases, e.g., Network Slicing, RAN QoE optimization, IAB network, etc. [RAN2, RAN3]:
  - Study and evaluate data collection, input/output data of model training and inference, and their standardization impacts
  - 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 between AI functions

intel®