3GPP TSG RAN #102

Edinburgh, UK, December 11 – 15, 2023

RP-233324

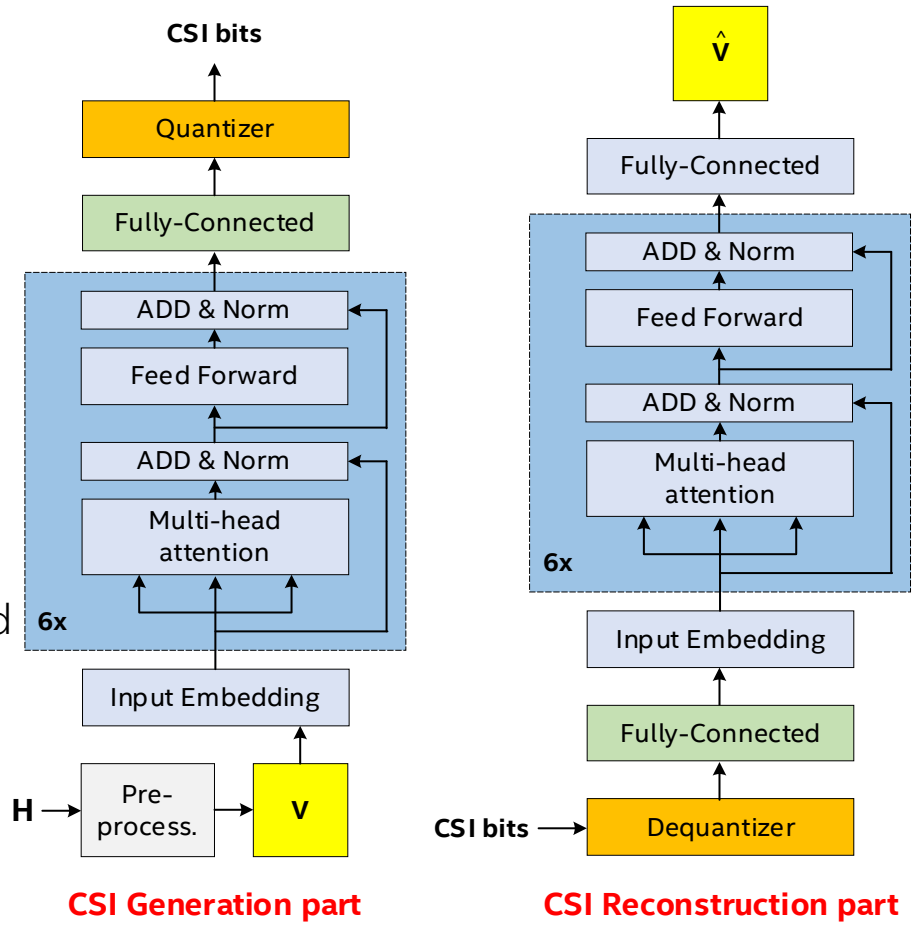
Discussion on Rel-19 AI/ML for Air Interface

Agenda Item:	9.1.1.1
Source:	Intel Corporation
Document for:	Discussion

intel®

Introduction – Rel-18 status

- Rel-18 AI-ML for air-interface SI considered the following
  - Terminology/framework/LCM for AI/ML
  - 3 use-cases: CSI, beam-management, positioning
  - CSI compression use-case based on 2-sided AI-ML model – CSI Generation part and CSI Reconstruction part (shown in figure)
- Overall observations in Rel-18
  - A lot of enthusiasm and evaluation results using state-of-the art AI/ML models (often with prohibitive complexity)
  - More convergence and time is needed for normative specification work in many cases
  - No consensus on whether a NW trained model can be downloaded and deployed at a UE in Rel-19 time-frame
  - Unclear whether proprietary format models can be beneficial for large scale deployment and usage
  - Concerns on disclosing model structure for training
  - RAN4 feedback (performance requirements) not available yet

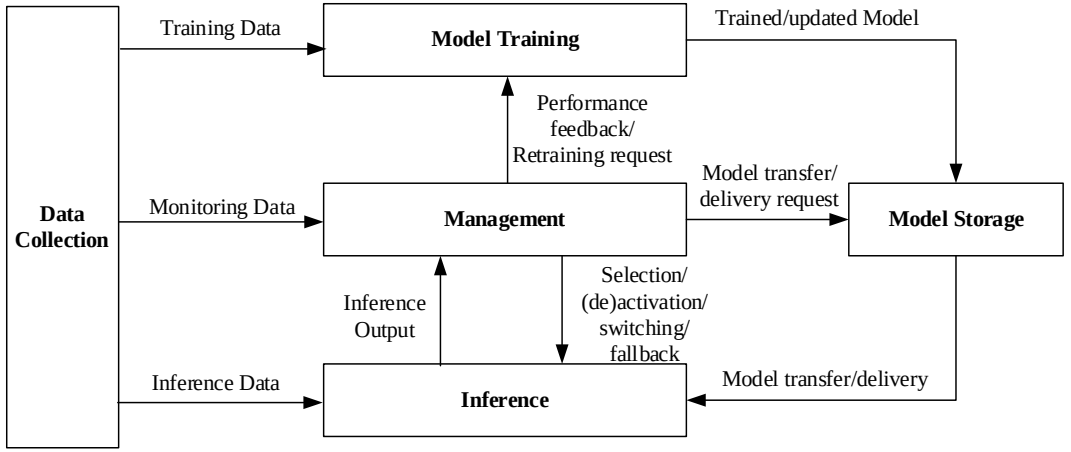


General Aspects 1/2

Justification

- 3GPP has studied the support of AI/ML framework
- Data collection: RAN2 agreed some basic assumptions on data collection latency requirement, RRC state, termination
- Model usage: Model control and related signaling needs further study in RAN2
- Collaboration levels: little convergence on the specification details that are required between gNB and UE for model storage, transfer and delivery
- Functional framework and functionality mapping: agreed in RAN2

- Model transfer/delivery: RAN2 has identified 7 solutions, where model transfer can be done over CP or UP. Further down-selection is possible according to the mapping between functionality and physical entity



General aspects 2/2

Objectives

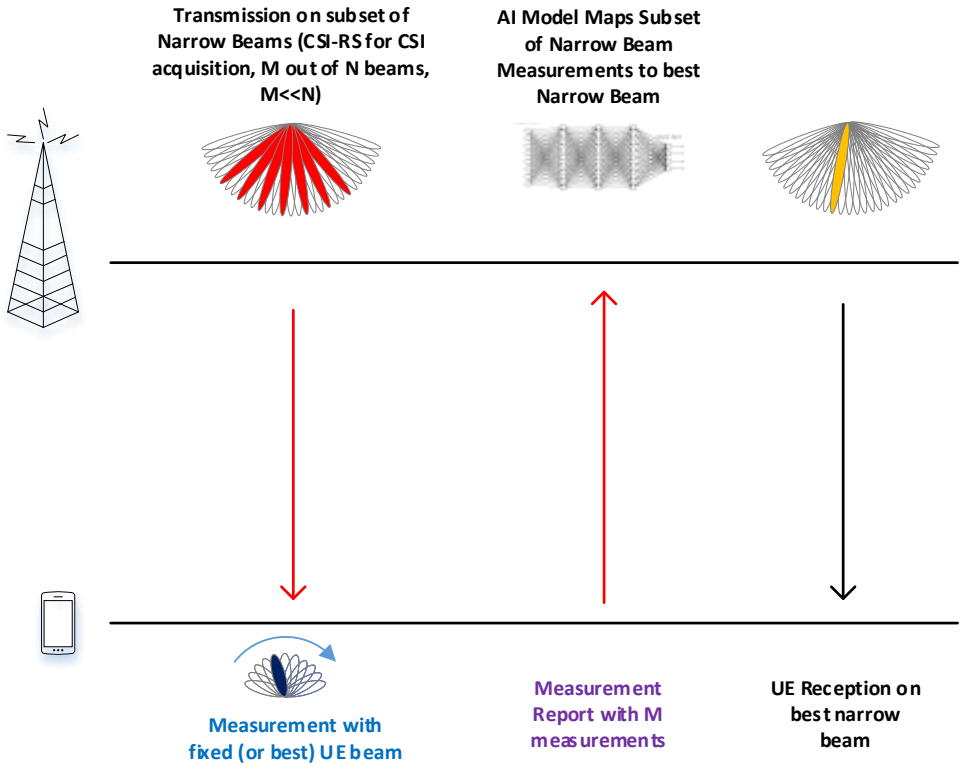
- Consider specifications for
  - NW side or UE side models without model transfer
  - Functionality based LCM without model identification
    - Based on UE capability framework
  - Signaling/procedures necessary for
    - Ensuring consistency between training and inference (additional conditions)
    - Data collection for training, inference, monitoring (assistance information)
    - Fallback operation (monitoring)

- Continue study for both single sided and two-sided models
  - Procedures for
    - model identification (online and offline), model selection, activation, deactivation, switching, and fallback operation
    - validation and monitoring of inactive models
  - Model delivery/transfer for all collaboration levels

Beam Management 1/2

Justification

- Large number of analog TX beams at gNB and Rx beams at UE (with multiple panels) are common in higher frequency bands (above 28GHz)
- L1-RSRP measurement and reporting overhead needed for best Tx/Rx beam selection can be prohibitively large and incurs large latency
- AI/ML based Beam Prediction can reduce measurement RS overhead as well as reporting latency
- 3GPP studied two main sub-use cases in Rel-18 e.g., spatial and temporal domain beam prediction and simulations showed reasonably good gains for Tx beam prediction as well as Tx-Rx beam pair link prediction



Beam management 2/2

Objectives

- Specifications for spatial domain and temporal domain DL Tx beam prediction with both UE and NW side models
- Study and, if justified, specify Tx-Rx beam-pair prediction with NW-side and UE-side models
  - Framework for L1 measurement/reporting and beam indication for NW and UE ensuring proprietary analog beamforming information is not shared across the air-interface
- Specification for enhanced L1 measurement/reporting for data collection and reference signal configurations for LCM for both UE sided and NW sided models
  - L1 measurement and beam reporting enhancements for training, inference and model monitoring
    - Reference signal configuration and triggering of reference signal transmissions from NW and/or UE for LCM purposes

Study for CSI 1/2

Justification

- Trade-off between performance and complexity (compute, storage)/overhead
  - Reported complexity for each part of a 2-sided model ~ 10M – 1000M FLOPs compared to non-AI-ML CSI complexity < 1M FLOPs
  - Study of time-dimension compression may not resolve the complexity concern, while adding significant TU burden (performance evaluations required for many aspects, including gain over benchmark, generalization, scalability, training, etc.)
- Issues related to inter-vendor training collaboration
  - Type 2 and Type 3 training does not require disclosure of model but there is no consensus on aspects such as extendibility of a UE part or NW part of a model and flexibility to support site specific models
- Issues related to inter-vendor collaboration for model delivery/transfer
  - Model transfer/delivery is a mechanism to handle site specific models to address the complexity concern
  - Inter-vendor collaboration for model storage/transfer, proprietary model disclosure concerns

Study for CSI 2/2

Objectives for CSI compression

- Study trade-off between performance and complexity (storage, compute)/overhead
  - Study site specific models with complexity comparison with non-AI/ML
  - Conclusion on CQI/RI feedback
- Study inter-vendor training collaboration
  - Focus on resolving aspects of Type-1, Type-2, Type-3 training aspects such as extendibility, flexibility to support site specific models and allowing UE side and NW side to develop/update models separately

Objectives for CSI prediction

- Study specification impact considering re-use of CSI prediction framework from Rel-18 MIMO
  - CSI-RS, PMI codebook, CQI calculation, CSI prediction timing, etc.
- Study model performance monitoring based on intermediate metrics (e.g., SGCS) calculated from measured CSI-RS and predicted channel at the UE

intel®