vivo

3GPP TSG RAN Rel-19 workshop RP-233048

Edinburg, December 11th ~ 15th , 2023

Agenda Item: 9.1.1.1

Source: vivo

Title: Rel-19 items for AI/ML over Air interface

Document for: Discussion

Observation from offline discussion in RANP #101

- **Observations based on the input in RAN#101**
  - Consider support single sided models for,
    - Positioning, further discuss whether to support all 5 sub use cases.
    - Beam management for DL Tx prediction in both UE-sided and NW sided model,
  - Consider support of single sided model for CSI prediction, if it is recommended to proceed with normative work taken all WG’s outcome in Q4-23 into account
  - Consider support two-sided model for CSI compression, if it is recommended to proceed with normative work taken all WG’s outcome in Q4-23 into account
  - Consider support the necessary/recommended LCM components for selected sub use cases
  - RAN4 requirements and test cases for AI/ML enabled features

Summary of Topics and views

- **One sided use case:**
  - R19 normative work on beam and positioning use cases need further study on the appropriate approach for additional condition alignment between NW-side and UE-side
  - Positioning: Support all 5 use cases of AI/ML based positioning for Rel-19 AI/ML WI
  - Beam: Support both BM-Case1 and BM-Case2 of AI/ML based beam management for Rel-19 AI/ML WI
  - CSI prediction: Study CSI prediction and check at 2024 September for normative work
- **Two sided use case:**
  - Study two-sided use cases and check at 2024 September for normative work, including
    - Achieving better performance/complexity tradeoff for CSI compression, including site/scenario-specific model, CSI temporal compression
    - Conclude training types of two-sided, including aspects related to model transfer and dataset delivery
    - Open study of new use cases: DMRS with joint sequence and estimation design, PA with joint transmitter and receiver design
- **Aspects in General Framework:**
  - Study model level LCM and check at 2024 September for normative work, including
    - Model transfer: Conclude and down-select the feasible options
    - Model identification: Conclude on necessity and details of model identification types
  - Data collection:
    - Trigger SA WGs to further study UE-side data collection with standardized content
    - NW-side data collection for model training should consider UE memory, processing power, energy consumption, signaling overhead and user privacy
- **RAN4 arrangement:**
  - Further study testability of two-sided and one-sided use cases from 2024 Q1 with checkpoint on 2024 September for new use cases
  - Specify core requirement and performance requirements for supported use cases and LCM components
- **Coordination with SA is needed for Rel-19 work**

One sided use cases

Additional condition alignment

- 4 options to align NW-side additional conditions are identified but have not been fully studied:
  - Model identification to achieve alignment on the NW-side additional condition between NW-side and UE-side
  - Model training at NW and transfer to UE, where the model has been trained under the additional condition
  - Information and/or indication on NW-side additional conditions is provided to UE
  - Consistency assisted by monitoring (by UE and/or NW, the performance of UE-side candidate models/functionalities to select a model/functionality)
- Possible additional conditions for use cases
  - BM-specific conditions/additional conditions including at least: information regarding model inference, set A / Set B configuration, performance monitoring, data collection, assistance information
  - Additional conditions may be used by AI/ML based positioning Case 1 and Case 2a (model is at UE-side)

- **Proposal: R19 normative work on beam and positioning use cases need further study on the appropriate approach for additional condition alignment between NW-side and UE-side.**

One sided use cases

Positioning accuracy enhancements

- In Rel-18 study, it is recommended to proceed with normative work for AI/ML based positioning.
- For all five positioning cases (Case 1/2a/2b/3a/3b), RAN1 has not considered prioritization.
- It is recommended to specify necessary measurement, signalling and procedure to facilitate training, inference, monitoring and/or other LCM operations for both direct AI/ML positioning and AI/ML assisted positioning:
  - Specify necessary signalling of data collection; investigate the necessity of other information for supporting data collection, and if needed, specify during normative work
  - Investigate on the necessity and signalling details of measurement enhancements, and if needed, specify during normative work
  - Investigate on the necessity and signalling details of monitoring method(s), and if needed, specify during normative work

- **Proposal: Support all 5 use cases of AI/ML based positioning for Rel-19 AI/ML WI.**

One sided use cases

Beam management

- In Rel-18 study, for AI-based beam management, from RAN1 perspective, at least the following are recommended for normative work:
  - Both BM-Case1 and BM-Case2:
    - BM-Case1: Spatial-domain DL Tx beam prediction for Set A of beams based on measurement results of Set B of beams
    - BM-Case2: Temporal DL Tx beam prediction for Set A of beams based on the historic measurement results of Set B of beams
  - DL Tx beam prediction for both UE-sided model and NW-sided model
  - Necessary signalling/mechanism(s) to facilitate data collection, model inference, and performance monitoring for both UE-sided model and NW-sided model
  - Signalling/mechanism(s) to facilitate necessary LCM operations via 3GPP signalling for UE-sided model

- **Proposal: Support both BM-Case1 and BM-Case2 of AI/ML based beam management for Rel-19 AI/ML WI.**

One sided use cases

CSI prediction

- From RAN1 perspective, there is no consensus on the recommendation of CSI prediction for normative work, and the reason is
  - Lack of results on the performance gain over non-AI/ML based approach and associated complexity.
- Rel-18 study of CSI prediction mainly includes
  - The performance and potential specification impact were studied for AI/ML based UE side CSI prediction sub use case.
  - Evaluations includes performance compared to baseline, model input/output type, generalization over UE speed, etc. Some aspects are studied but lack observations, including scalability over various configurations and generalization over other scenarios and approach of fine tuning. Performance monitoring accuracy has not been evaluated.
  - Potential specification impact on data collection and performance monitoring are discussed. Limited specification aspects were considered.

- **Proposal: Study CSI prediction and check at 2024 September for normative work, including**
  - **Sufficient simulation results**
  - **Specification impact especially for model level LCM**

Two sided use cases

Two-sided use cases

- From RAN1 perspective, there is no consensus on the recommendation of CSI compression for normative work due to trade-off between performance and complexity/overhead and issues related to inter-vendor training collaboration
  - R19 can further study better performance/complexity tradeoff, including
    - Site/scenario-specific model: model trained on specific scenarios (e.g., cell-level area) provides more Tput gain (about 30%) compared to eType II codebook, which is much higher than that reported by companies in R18 studies (about 8%~10%)
    - Temporal domain CSI compression: provides additional Tput gain compared to CSI compression studied in R18
    - Above two methods can be combined together for higher performance gain
  - R19 can further study and conclude the feasible inter-vendor training collaboration types, including model transfer (Type 1), dataset delivery (Type 3)
- New use cases may be also studied in R19
  - E.g., DMRS with joint sequence and estimation design, PA with joint transmitter and receiver design

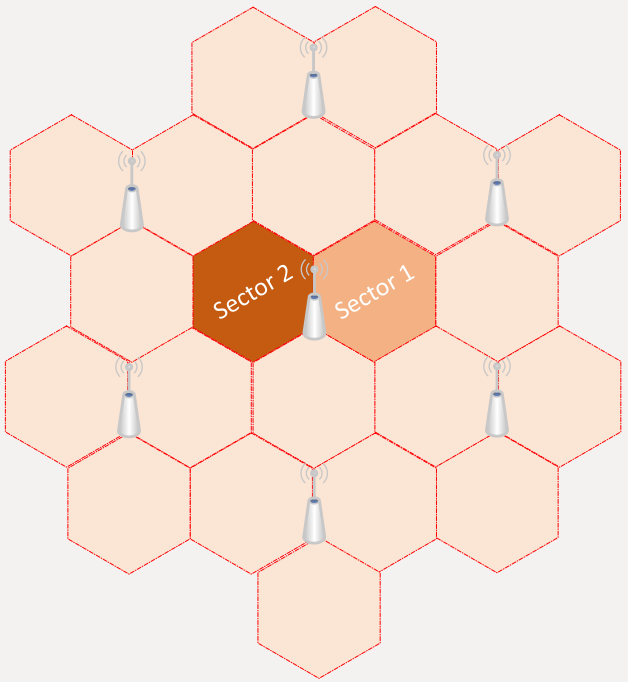
- **Proposal: Study two-sided use cases and check at 2024 September for normative work, including**
  - **Achieve better performance/complexity tradeoff for CSI compression, including site/scenario-specific model, CSI temporal compression**
  - **Conclude training types of two-sided, Including model transfer, dataset delivery**
  - **Open study of new use cases: DMRS with joint sequence and estimation design, PA with joint transmitter and receiver design**

Two sided use cases

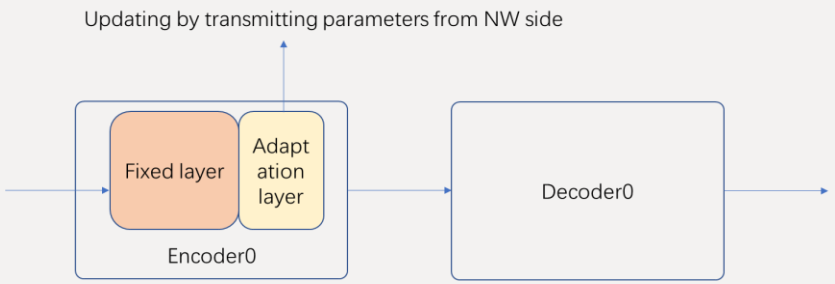
Site/scenario-specific model

- Initial results on site/scenario-specific model (e.g., cell-level area)

	SGCS on Sector 1 (heavy LOS)	SGCS on Sector 2 (mixed LOS and NLOS)
Legacy R16 codebook	0.8813	0.6818
MLP encoder	0.9741 (+10.53%)	0.8954 (+31.33%)
Transformer encoder	0.9779 (+10.96%)	0.9121 (+33.78%)
General model with tunable encoder and decoder	0.9705 (+10.12%)	0.8969 (+31.55%)
General model with tunable adaptation layer and decoder	0.9617 (+9.12%)	0.8538 (+25.23%)



- Site/scenario-specific model provides more gains (about SGCS 34% for sector 2) compared to what is reported by companies in R18 studies (2.6%~8.8%)
- Simple model (e.g. MLP), also harvests major performance gain of complex model (e.g. Transformer) for site/scenario-specific model
- Tunable adaptation layer and decoder provides higher gains for cell-site specific models, although suffers slight performance loss compared to tunable encoder and decoder

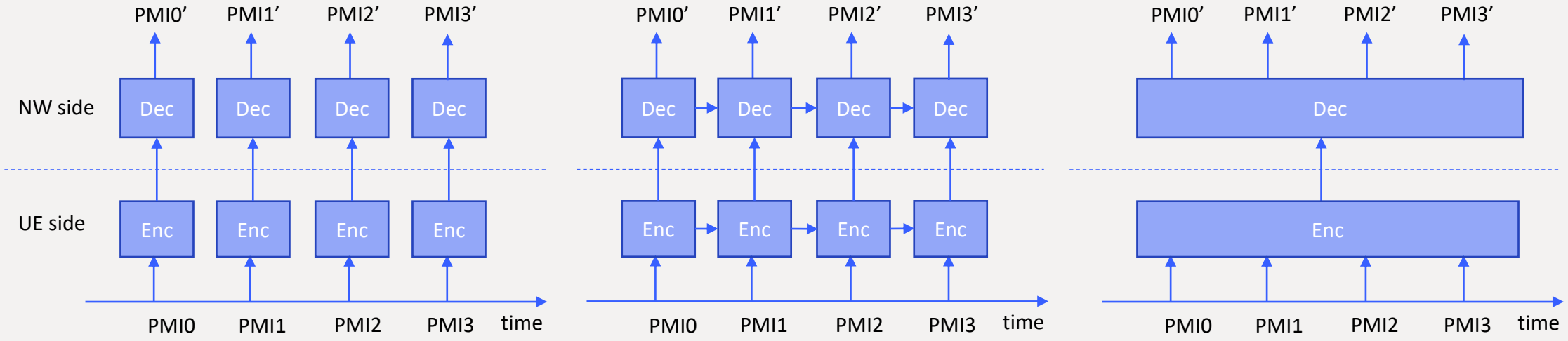


Simulation parameters: Dense urban with spatial consistancy, UMa 38.901, carrier frequency 4GHz, subcarrier spacing 30KHz, 13 subbands (4RBs/subband), 32 gNB antenna ([Mg Ng M N P; Mp Np] = [1 1 8 8 2; 2 8]), 2 UE antenna ([Mg Ng M N P; Mp Np] = [1 1 1 1 2; 1 1]), 100% outdoor UE, 30km/h, layer0 considered. Overhead is 64bits. General model is trained by 21 sectors without spatial consistency.

Two sided use cases

T-S-F compression

- Principle: exploiting temporal correlation in consecutive CSIs to further improve the performance gains



Benchmark: s-f domain CSI compression in R18

AR T-S-F: Auto-regressive t-s-f domain CSI compression

J T-S-F: Joint t-s-f domain CSI compression

	Legacy R16 codebook	Observation from Rel-18 study 64*4	AR T-S-F with payload 64*4	J T-S-F with payload 64*4
Average SGCS	100%	+2.6%~8.8%	+14.07%	+17.21%

- Exploiting temporal domain correlation could further provide relative SGCS gain on top of s-f domain compression
- Overhead could be further reduced by ~60% compared with s-f model, and the overall overhead reduction compared with R16 CB improves from ~30% to ~70%

Simulation parameters: Urban Macro, with 500 ISD, carrier frequency 3GHz, subcarrier spacing 30KHz, 13 subbands (4RBs/subband), 32 gNB antenna ([Mg Ng M N P; Mp Np] = [1 1 2 8 2; 2 8]), 2 UE antenna ([Mg Ng M N P; Mp Np] = [1 1 1 1 2; 1 1]), 100% outdoor UE, 30km/h, rank1 layer0 considered.

Aspects in general framework

Model transfer and model identification

- Model transfer:
  - It is concluded in RAN1 that model transfer/delivery is beneficial at least in the following cases:
    - Model transfer/delivery can support flexible scenario/configuration specific (including site-specific configuration/channel conditions) models, which may provide performance benefits in some studied use cases (i.e., when a single model cannot generalize well to multiple scenarios/configurations/sites).
    - Model transfer/delivery without offline compiling and/or testing (that can update parameter with known model structure), may have benefit at least in terms of shorter model parameter update timescale.
    - Model transfer/delivery can ensure consistency between training and inference regarding NW-side additional conditions of UE-side models, when model is trained at NW and transferred to UE, where the model has been trained under the additional condition.
    - Model transfer can be used one-sided model with model inference at UE but data collection at e.g. gNB.
  - RAN2 has studied all model transfer/delivery cases and captures the RAN specification potential impacts into TR. RAN2 captures that both Reactive model transfer/delivery and Proactive model transfer/delivery can be considered in normative phase.
  - Rel-19 needs to conclude and down-select feasible options based on further study.
- Model identification:
  - Companies have different understandings on details of model identification Types A, B1, B2

- **Proposal: Study model level LCM and check at 2024 September for normative work, including**
  - **Model transfer: Conclude and down-select the feasible options**
  - **Model identification: Conclude on necessity and details of model identification types**

Aspects in General framework

UE-side Data collection

- For UE-side data collection, the following approaches were discussed:
  - 1. UE collects and directly transfers training data to the Over-The-Top (OTT) server;
    - 1a) OTT (3GPP transparent)
    - 1b) OTT (non-3GPP transparent)
  - 2. UE collects training data and transfers it to Core Network. Core Network transfers the training data to the OTT server.
  - 3. UE collects training data and transfers it to OAM. OAM transfers the needed data to the OTT server.
- The majority think solutions 1b/2/3 are out of RAN2 scope, thus RAN2 did not study or analyze these approaches and did not agree to the requirements or recommendations. In this case, RAN2 should trigger SA WGs to further study UE-side data collection mechanism.
- In addition, the study should focus on standardized/specified data content.

- **Proposal: Trigger SA WGs to further study UE-side data collection with standardized content.**

Aspects in General framework

NW-side Data collection

- During SI phase, RAN2 concluded on a set of general principles of NW-side data collection for model training, including:
  - UE to support data logging,
  - UE to report the collected data periodically, event-based, and on-demand,
  - The UE memory, processing power, energy consumption, signaling overhead should be considered.
- Furthermore, regarding the use cases, the following mechanisms are considered:
  - For CSI and beam management use cases, the data collection for training can be gNB-/OAM-centric, which may introduce impact on L3 signaling and MDT. From our understanding, the data collection for training and inference can be UE-agnostic, thus L1 signaling can also be considered.
  - For positioning use cases, the data collection for training can be LMF-centric, and the LPP or NRPPa are assumed to be applicable.
- Similar to AI for NG-RAN, it is a high-level principle that user data privacy needs to be preserved.

- **Proposal: NW-side data collection for model training should consider UE memory, processing power, energy consumption, signaling overhead and user privacy.**

RAN4 Arrangement

- Beam management and positioning use cases
  - In R18 study, possible requirement metrics were studied for both beam management and positioning;
  - R19 could study the testability of one-sided use cases from 2024 Q1, and start the normative work to specify RRM core requirements for both beam management and positioning use cases;
- RRM core requirements for LCM
  - In R18 study, possible requirement metrics were studied for LCM framework;
  - R19 could start the normative work to specify RRM core requirements for functionality level LCM and for model level LCM based on RAN1 progress, if necessary;
- CSI compression use case
  - Testability of two-sided need further study, e.g. pros/cons analysis and test procedure of four options of test decoder;

- **Proposal: For R19 RAN4,**
  - **Further study testability of two-sided and one-sided use cases from 2024 Q1 with checkpoint on 2024 September for new use cases;**
  - **Specify RRM core requirements for beam management and positioning use cases;**
  - **Specify RRM core requirements for functionality level LCM and for model level LCM based on RAN1 progress, if necessary;**
  - **Specify RRM performance requirements and test cases for all supported use cases;**

Coordination with SA

- At least the following areas are identified with SA impact for the three studied use cases:
  - AI + Positioning procedures and signaling
    - AI/ML based positioning is expected to be supported in Rel-19. AI/ML based positioning support is heavily related to SA work.
  - Model identification procedures and signaling
    - Model identification procedure is expected to have SA impact.
  - Model transfer/Data collection related issues;
    - Model transfer/delivery can support flexible scenario/configuration specific model, and can ensure consistency between training and inference regarding NW-side additional conditions of UE-side models.
    - Data collection solutions are expected to have SA impacts. Detailed content of data collection can be sent to SA based on RAN1 LS reply in R1-2308730 and R1-2310681.

- **Proposal: RAN shall inform SA on the RAN1 recommendation of AI/ML based positioning, RAN1 agreements of model transfer/delivery, data collection, model identification and also RAN planning on R19 WI plan for above aspects if agreed.**

vivo

THANK YOU.
谢谢。