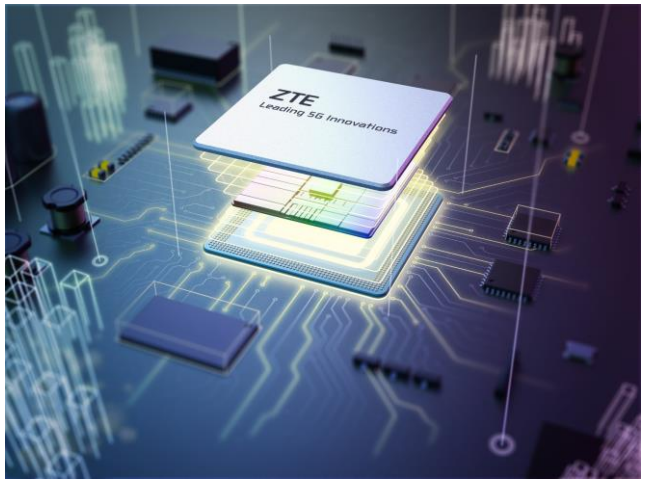
3GPP TSG RAN Meeting #102

Edinburgh, Scotland, December 11-15, 2023

Agenda item: 9.1.1.1

TDoc number: RP-233608

ZTE



Views on Rel-19 AI/ML for Air Interface

Agenda

1. Summary of Rel-18 AI/ML PHY
2. Analysis and Proposal
3. Summary of draft objectives

Summary of Rel-18 AI/ML PHY

The Rel-18 study on AI/ML for NR Air Interface was approved (RP-221348)

- RAN1 study was planned to be finished by 2023 Q3, and RAN2/4 study was planned to be finished by 2023 Q4
- The RAN1 study on AI/ML for NR Air Interface was delayed by one quarter. RAN1 focused on the following in 2023 Q4
  - General aspects of AI/ML framework
  - Other aspects on AI/ML for CSI feedback enhancement
  - Finalization of TR conclusions/recommendations

Use cases and sub-use cases for Rel-18 study on AI/ML for NR Air Interface

- CSI feedback enhancement
  - Spatial-frequency domain CSI compression using two-sided AI model
  - Time domain CSI prediction using UE sided model
- Beam management
  - BM-Case1: Spatial-domain Downlink beam prediction
  - BM-Case2: Temporal Downlink beam prediction
- Positioning accuracy enhancements
  - Direct AI/ML positioning
  - AI/ML assisted positioning

© ZTE All rights reserved

Summary of Rel-18 AI/ML PHY

Overall progress on the Rel-18 study on AI/ML for NR Air Interface (RAN1)

Use case	Sub-use case	Overall progress
CSI	Spatial-frequency domain CSI compression	1. Simulation results show **limited performance** gain 2. Potential solutions/methods for supporting CSI compression were identified but not concluded 3. LCM for two-sided model was studied but not concluded 4. **No consensus in RAN1 to recommend CSI compression for normative work**
	Time domain CSI prediction	1. Simulation results show **limited performance** gain 2. Potential solutions/methods for supporting CSI prediction were identified but not concluded 3. LCM for one-sided model follows what defined in the AI/ML framework agenda **4. No consensus in RAN1 to recommend time domain CSI prediction for normative work**
Beam	Spatial-domain Downlink beam prediction	1. Simulation results show moderate ~ high beam prediction gain 2. Potential solutions/methods for support AI/ML beam prediction were identified 3. LCM for one-sided model follows what defined in the AI/ML framework agenda 4. **Both sub-use cases are recommended for normative work by RAN1**
	Temporal Downlink beam prediction	
Position	Direct AI/ML positioning	1. Simulation results show significant position accuracy gain 2. Potential solutions/methods for support AI/ML position were identified 3. LCM for one-sided model follows what defined in the AI/ML framework agenda 4. **Both sub-use cases are recommended for normative work by RAN1**
	AI/ML assisted positioning	

© ZTE All rights reserved

Summary of Rel-18 AI/ML PHY

Overall progress on the Rel-18 study on AI/ML for NR Air Interface (RAN2)

Study Area	Overall progress
Model ID and Meta Info	1: Model ID is assumed to be unique 'globally' from RAN2 perspective. 2: Model ID can be used to identify a model from RAN2 perspective 3: The meta info is concluded to be used for model control/management from RAN2 perspective, but it is still FFS how to use.
Model Transfer/delivery	1: Up to 8 solutions on the table are studied, but **no recommendation is made**. 2: The mapping relationship between solutions and use cases are initially concluded, but **no recommendation is made**. 3: The current situation of the specification and the potential specification efforts to support each solution are concluded, but **no recommendation is made**.
Data Collection	1. There are 8 candidate solutions proposed for NW sided data collection in RAN2. 2: Immediate MDT with some potential enhancements is recommended as a baseline solution for NW sided data collection for model training. 3: There is **no any requirements and/or recommendation for UE sided data collection** for model training is concluded in RAN2.
Functionality mapping	1: Per use case, the mapping relationship between model training/transfer/management/control/inference and logical entities are concluded in RAN2. But the logical entities, such as OAM, CN, are considered as out of RAN2 scope in some mapping relationship.
Capability/Conditions	1: RAN2 concludes the legacy UE capability framework serve as the baseline to report UE's supported AI/ML-enabled Feature/FG 2:RAN2 concludes the UAI can be a baseline solution for UE to report the additional conditions.

© ZTE All rights reserved ZTE

Summary of Rel-18 AI/ML PHY

Overall progress on the Rel-18 study on AI/ML for NR Air Interface (RAN4)

Study Area	Overall progress
Testing goals	For testing goals, Option 1 and/or Option 2 are studied **without conclusion/recommendation** • Option 1 (Model-level): The testing goal is to verify whether a specific AI/ML model (if model identification is possible)/functionality can be conducted in a proper way. • Option 2 (Performance-level): The testing goal is to verify whether the minimum performance gain of AI/ML model (if model identification is possible) /functionality/feature can be achieved for a static scenario/configuration.
Data Collection	Companies discussed the following methods for data collection for testing but **without any agreements/conclusion** • Dataset based on TR 38.901, e.g. UMa channel, UMi channel, CDL channel, “legacy approach”, etc. • Field dataset (data collected directly from field measurements) • TE generates dataset for test based on assumptions/parameters defined by RAN4 (e.g. by defining some rules/function to generate data)
Generalization	RAN4 agreed the following two methods for verification of generalization with multiple open issues and without any recommendation. • RAN4 may define multiple tests with different conditions. In each of the test, TE configures the same specified UE configuration, and therefore the same specified UE configuration is tested under different conditions to verify its generalizability. • The goals of generalization test are to verify whether the minimum level of performance of AI/ML functionality/model can be achieved/maintain under the identified scenarios and/or configurations, while the performance won’t be significantly degraded in other scenarios and/or configurations.
Reference block for test	RAN4 discussed the **general structure** of one-sided and two-sided models based on the RAN2’s AI/ML framework. For two-sided model, the pros and cons and the encoder/decoder are clarified with multiple detailed open issues.
KPI for use-cases	Different KPIs for three different use cases are discussed **without any conclusion/recommendation**.

© ZTE All rights reserved ZTE

Agenda

1. Summary of Rel-18 AI/ML PHY
2. **Analysis and Proposal**
3. Summary of draft objectives

Analysis and Proposal: Framework

Framework studied in RAN1

- The following aspects have been studied for the general framework of AI/ML over air interface for one-sided models and two-sided models (RAN1)
  - Various Network-UE Collaboration Levels
  - Functionality-based LCM and model-ID-based LCM
  - Functionality/model selection, activation, deactivation, switching, fallback
  - Functionality identification and model identification
  - Data collection
  - Performance monitoring
  - Various model identification Types and their use cases
  - Reporting of applicable functionalities/models
  - Method(s) to ensure consistency between training and inference regarding NW-side additional conditions (if identified) for inference at UE
  - Model delivery/transfer and analysis of various model delivery/transfer Cases
  - Note: The above studied aspects for General Framework can be considered for developing/specifying AI/ML use cases and common framework (if needed for some aspects) across AI/ML use cases.

© ZTE All rights reserved

Analysis and Proposal: Framework

Functionality-based LCM and Model-ID based LCM

- At the beginning of Rel-18 study, two LCMs are identified
  - **Functionality-based LCM** operates based on configuration(s) of AI/ML-enabled Feature/FG or specific configurations of an AI/ML-enabled Feature/FG
  - **Model-ID based LCM** operates based on identified models, where a model may be associated with specific configurations/conditions associated with UE capability of an AI/ML-enabled Feature/FG and additional conditions
- Later on, a unified framework is identified
  - Model-ID, if needed, can be used in a Functionality (defined in functionality-based LCM) for LCM operations
- Summary of comparison
  - Functionality-based LCM operates based on the legacy UE feature and RRC configuration framework, which is more implementation friendly for both base station and UE.
  - Model-ID based LCM operates on model level, which extremely increases the implementation burden of base station to control/manage the different models for different use cases from different vendors
  - Model-ID based LCM can be applied for two-sided models for model pairing

Proposal 1: Regarding LCM for Rel-19 AI/ML framework, support Functionality-based LCM for one-sided model.

© ZTE All rights reserved

Analysis and Proposal: Framework

Model delivery/transfer

- Different cases
  - Case y: Model delivery over-the-top
  - Case z: Model transfer over-the-air
- Summary of Pros and Cons
  - Model delivery/transfer to UE may be beneficial to handle scenario/configuration specific (including site-specific configuration/channel conditions) models (i.e., when a single model cannot generalize well to multiple scenarios/configurations/sites), to reduce the device storage requirement.
  - Model storage at the 3gpp network, compared to storing the model outside the 3gpp network, may come with 3gpp network side burden on model maintenance/storage.
  - Proprietary design disclosure concern may arise from model training and/or model storage at the network side compared to other cases (such as case y with UE side training) which does not have such issue.
- Feasibility of model deliver/transfer has NOT been confirmed

Case	Model delivery/transfer	Model storage location	Training location
y	model delivery (if needed) over-the-top.	Outside 3gpp Network	UE-side / NW-side / neutral site
z1	model transfer in proprietary format.	3GPP Network	UE-side / neutral site
z2	model transfer in proprietary format.	3GPP Network	NW-side
z3	model transfer in open format.	3GPP Network	UE-side / neutral site
z4	model transfer in open format of a *known model structure* at UE, i.e., an exact model structure as has been previously identified between NW and UE and for which the UE has explicitly indicated its support.	3GPP Network	NW-side
z5	model transfer in open format of *an unknown model structure* at UE, i.e., any other model structure not covered in z4, including any model structure that is only partially known.	3GPP Network	NW-side

Proposal 2: For one-sided model, implementation-based solution(s) can be used for model delivery/transfer.

© ZTE All rights reserved

Analysis and Proposal: Framework

Additional conditions discussed in RAN1

- Definition and motivation of additional conditions
  - For an AI/ML-enabled feature/FG, additional conditions refer to any aspects that are assumed for the training of the model but are not a part of UE capability for the AI/ML-enabled feature/FG.
  - Additional conditions can be divided into two categories: **NW-side additional conditions** and **UE-side additional conditions**.
  - For inference for UE-side models, it may be beneficial to ensure consistency between training and inference regarding NW-side additional conditions for some use cases based on the simulation results
- Detailed solutions to guarantee consistent additional conditions during training and inference
  - Opt1: Model identification to achieve alignment on the NW-side additional condition between NW-side and UE-side
  - Opt2: Model training at NW and transfer to UE, where the model has been trained under the additional condition
  - Opt3: Information and/or indication on NW-side additional conditions is provided to UE
  - Opt4: Consistency assisted by monitoring (by UE and/or NW, the performance of UE-side candidate models/functionalities to select a model/functionality)
- Summary of RAN1 discussion on additional conditions
  - Currently, no valid additional condition has been agreed for all use cases.
  - Additional conditions may disclose the implementation proprietary of base station
  - Among the 4 options above, only Opt4 has less/no specification impact and less possibility of disclosing implementation proprietary

© ZTE All rights reserved

Analysis and Proposal: Framework

Additional conditions discussed in RAN2

- RAN2 mainly discussed UE-side additional conditions.
  - For additional condition reporting, the existing capability reporting framework cannot be used. To report these conditions (if needed), UAI can be used as an example. This can be defined and discussed in normative phase.
- Information on UE-side additional conditions (e.g., UE speed or UE Rx beam pattern) is beneficial for network-side model or network management/control of UE-side model

Summary of discussion on additional conditions in RAN1 and RAN2

- RAN1 discussed both NW-side additional conditions and UE-side additional conditions, but progressed mainly on NW-side additional conditions.
- RAN2 mainly discussed and progressed on UE-side additional conditions.
- Both NW-side additional conditions and UE-side additional conditions have its pros and cons. Joint discussion on both of them is suggested from our perspective.

Proposal 3: Additional condition (NW-side additional conditions and UE-side additional conditions) is not specified explicitly in Rel-19. If necessary, additional condition can be handled by performance monitoring.

© ZTE All rights reserved

Analysis and Proposal: Framework

Data collection at UE side for model training

- No recommendation and/or requirements for data collection for training UE models are concluded in RAN2.
- The UE collecting data transfer from UE to UE side OTT server can be a implementation-based solution

Proposal 4: No more discussion on UE collecting data transfer to UE side OTT server for UE model training in Rel-19.

Model Id in RAN2/RAN1

- Model ID is assumed to be globally unique from RAN2 perspective while model ID is assumed to be a local ID from RAN1 perspective.
- The generalization of the model is not ideal in some use cases (e.g. positioning)
- The global model ID need more specification effort in SA2/CT1 rather than in RAN.

Proposal 5: If model identification is required, the model ID is considered as a local ID instead of a global ID in Rel-19.

© ZTE All rights reserved

Analysis and Proposal: CSI feedback enhancement

Simulation results/observations for CSI compression

- Max Rank 1/2: Most cases achieve less than 5% UPT gain
- Max Rank 4: Most cases achieve around 10% UPT gain
- The performance may further decrease considering the generalization/scalability, multi-vendor joint training, separate training, etc.

Mean UPT gain over benchmark for FTP traffic (Median value)

	RU	Overhead A	Overhead B	Overhead C
Max Rank 1	RU<=39%	1.25% (6 sources)	0.6% (6 sources)	0.65% (4 sources)
	RU 40%-69%	2.4% (5 sources)	1.5% (4 sources)	2.5% (7 sources)
	RU>=70%	3.6% (9 sources)	2.43% (8 sources)	1.85% (8 sources)
Max Rank 2	RU<=39%	2% (7 sources)	2% (7 sources)	2% (8 sources)
	RU 40%-69%	4.3% (8 sources)	4.26% (8 sources)	4.28% (10 sources)
	RU>=70%	11% (11 sources)	6.8% (11 sources)	5.5% (10 sources)
Max Rank 4	RU<=39%	6% (3 sources)	6% (1 source)	-2% (2 sources)
	RU 40%-69%	11% (3 sources)	9.02% (2 sources)	5.1% (3 sources)
	RU>=70%	14.89% (3 sources)	11.82% (2 sources)	6% (3 sources)

Aspects that are NOT concluded for CSI compression

- Methods to improve the performance gain over legacy CSI feedback
- Impact of CQI/RI determination particularly when maximum rank 4 is configured
- Evaluation on UE side proxy model and analysis on potential specification impact on supporting of UE side proxy model
- Further analysis on feasibility/necessity/benefits of each training collaboration type
- Conclude the necessity of supporting high resolution ground-truth format for NW side data collection and NW side performance monitoring
- Further analysis on dataset delivery

© ZTE All rights reserved

Analysis and Proposal: CSI feedback enhancement

Aspects that are studied for CSI prediction

- Evaluation has been performed to assess AI/ML based CSI prediction from various aspects: performance compared to baseline, model input/output type, generalization over UE speed, Scalability over various configurations, generalization over other scenarios and approach of fine tuning, etc.
- Potential specification impact has been identified from the following aspects
  - data collection
  - performance monitoring

Simulation results of CSI prediction under FTP traffic (Median value)

Mean UPT	Without spatial consistency	With spatial consistency
Benchmark#1 (nearest historical CSI)	4.9% (5 sources)	17.2% (3 sources)
Benchmark#2 (non-AI/ML based CSI prediction)	2.3% (2 sources)	0.7% (3 sources)

Proposal 6: AI/ML based CSI feedback enhancement (CSI compression and CSI prediction) needs further study in Rel-19 due to the following reasons.

1) Lack of performance gain
2) Multiple LCM issues and RAN1 specification impacts need further investigation
3) Lots of open issues for two-sided models in RAN4

© ZTE All rights reserved

Analysis and Proposal: Beam management

Simulation summary/observations for spatial/temporal domain beam prediction

- Basic performance for BM-Case1
  - For BM-Case1 when Set B is a subset of Set A or when Set B is different than Set A, AI/ML can achieve good performance with measurements of fixed Set B that is 1/4 or 1/8 of Set A of beam measured with best Rx beam for DL Tx beam prediction.
  - With 1/4 or 1/8 measurement/RS overhead, 96%~99% or 85%~98% of UE average throughput and 95%~97% or 70%~84% of UE 5%ile throughput of non-AI baseline option (exhaustive search over Set A beams) can be achieved according to the predicted beam from AI/ML.
- Basic performance for BM-Case2
  - For BM-Case2 when Set B is equal to Set A, AI/ML can achieve decent beam prediction accuracy with 1/5~1/2 measurement/RS overhead reduction without UE rotation. The longer the prediction time, the higher gain of beam prediction accuracy can be achieved by AI/ML.
  - For BM-Case2 when Set B is a subset of Set A, AI/ML can achieve good prediction accuracy with 1/2, 1/3, 1/4, 1/8 RS overhead in spatial domain, for the case Set B is fixed or variable with pre-configured patterns of beams with or without UE rotation.

Aspects that need to be enhanced

- Performance monitoring
  - Configuration/signaling from gNB to UE for measurement and/or reporting
  - Indication/request/report from UE to gNB for performance monitoring
  - Indication from NW to UE to do LCM operation
- Data collection
  - UE reporting to NW supported/preferred configurations of DL RS transmission
  - Signaling/configuration/measurement/report for data collection
  - Reporting overhead reduction via omission/selection/compression of collected data
- ……

© ZTE All rights reserved

Analysis and Proposal: Beam management

Aspects that need to be enhanced

- ……
- Model inference related
  - Indication of the associated Set A from network to UE
  - Beam indication from network for UE reception
  - Reporting information about measurements of multiple past time instances in one reporting instance
- L1 signaling
  - L1 signaling to report the beam(s) of one or multiple future time instances based on model output
  - L1 signaling to report predicted L1-RSRP or confidence/probability information related to the model output
  - L1 signaling to report the measurement results of more than 4 beams in one reporting instance
- LCM operations
  - BM-specific conditions/additional conditions for functionality(ies) and/or model(s)
  - Signaling/mechanism(s) to facilitate necessary LCM operations via 3GPP signaling

Proposal 7: AI/ML based beam management (BM-Case1 and BM-Case2) is included in Rel-19 AI/ML normative work, including performance monitoring, data collection, model inference related enhancement, L1 signaling, LCM operations, etc.

© ZTE All rights reserved

Analysis and Proposal: Positioning accuracy enhancements

Simulation summary/observations for positioning

- Evaluation results show that the enriched channel information (e.g., additional paths, PDP and CIR) can significantly increase the positioning performance in some scenarios (e.g., heavy NLos scenarios)
  - For example, in InF-DH with clutter parameter setting {60%, 6m, 2m}, AI/ML based positioning can achieve horizontal positioning accuracy of <1m at CDF=90%, as compared to >15m for conventional positioning method.
- Model generalization cannot be maintained across drops/clutter parameters/InF scenarios

Analysis of sub-use cases for AI/ML positioning accuracy enhancements

- Sub-use cases
  - Case 1: UE-based positioning with UE-side model, direct AI/ML or AI/ML assisted positioning
  - Case 2a: UE-assisted/LMF-based positioning with UE-side model, AI/ML assisted positioning
  - Case 2b: UE-assisted/LMF-based positioning with LMF-side model, direct AI/ML positioning
  - Case 3a: NG-RAN node assisted positioning with gNB-side model, AI/ML assisted positioning
  - Case 3b: NG-RAN node assisted positioning with LMF-side model, direct AI/ML positioning
- It can be up to UE/gNB implementation for UE/gNB-side model
- Prioritize enhancements to facilitate LMF-side model in Rel-19 for AI/ML based positioning

Proposal 8: Both direct AI/ML positioning and AI/ML assisted positioning are included in Rel-19 AI/ML normative work

- Prioritize enhancements to facilitate LMF-side model in Rel-19 for AI/ML based positioning

© ZTE All rights reserved

Agenda

1. Summary of Rel-18 AI/ML PHY
2. Analysis and Proposal
3. **Summary of draft objectives**

Summary of draft objectives

Summary of proposals

Proposal 1: Regarding LCM for Rel-19 AI/ML framework, support Functionality-based LCM for one-sided model.

Proposal 2: For one-sided model, implementation-based solution(s) can be used for model delivery/transfer.

Proposal 3: Additional condition (NW-side additional conditions and UE-side additional conditions) is not specified explicitly in Rel-19. If necessary, additional condition can be handled by performance monitoring.

Proposal 4: No more discussion on UE collecting data transfer to UE side OTT server for UE model training in Rel-19.

Proposal 5: If model identification is required, the model ID is considered as a local ID instead of a global ID in Rel-19.

Proposal 6: AI/ML based CSI feedback enhancement (CSI compression and CSI prediction) needs further study in Rel-19 due to the following reasons**.**

1) Lack of performance gain
2) Multiple LCM issues and RAN1 specification impacts need further investigation
3) Lots of open issues for two-sided models in RAN4

Proposal 7: AI/ML based beam management (BM-Case1 and BM-Case2) is included in Rel-19 AI/ML normative work, including performance monitoring, data collection, model inference related enhancement, L1 signaling, LCM operations, etc.

Proposal 8: Both direct AI/ML positioning and AI/ML assisted positioning are included in Rel-19 AI/ML normative work

- Prioritize enhancements to facilitate LMF-side model in Rel-19 for AI/ML based positioning

© ZTE All rights reserved ZTE

Summary of draft objectives

Summary of draft objectives

Framework

- Functionality-based LCM for one-sided model with compatibility to support the two-side model (if supported with check point)

AI/ML based beam management

- Support both BM-Case1 and BM-Case2
- Support performance monitoring, data collection, model inference related enhancement, L1 signaling, LCM operations for both BM-Case1 and BM-Case2

AI/ML positioning accuracy enhancements

- Support both direct AI/ML positioning and AI/ML assisted positioning
- Prioritize LMF-side model in Rel-19 for AI/ML based positioning
- Support signaling/report/feedback related enhancements, training data generation, performance monitoring, model inference related enhancement and LCM operations.

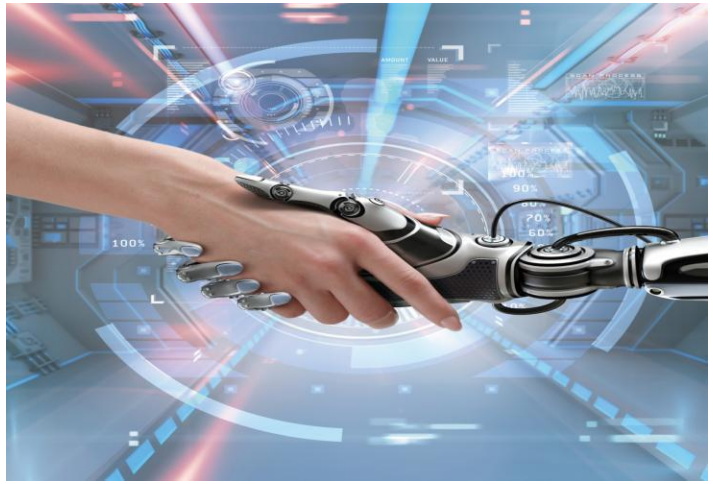
AI/ML CSI feedback enhancement

- Further study CSI compression based on two-sided models, including trade-off between performance and complexity/ overhead, issues related to inter-vendor training collaboration and other open issues.
- Further study CSI time-domain prediction, including solutions for improving performance gain over non-AI/ML based approach and associated complexity

RAN4 test/requirement part

© ZTE All rights reserved

ZTE



Thank You!