

**3GPP TSG RAN Rel-19 workshop**

**RWS-230063**

**Taipei, June 15 - 16, 2023**

**Agenda Item: 5**

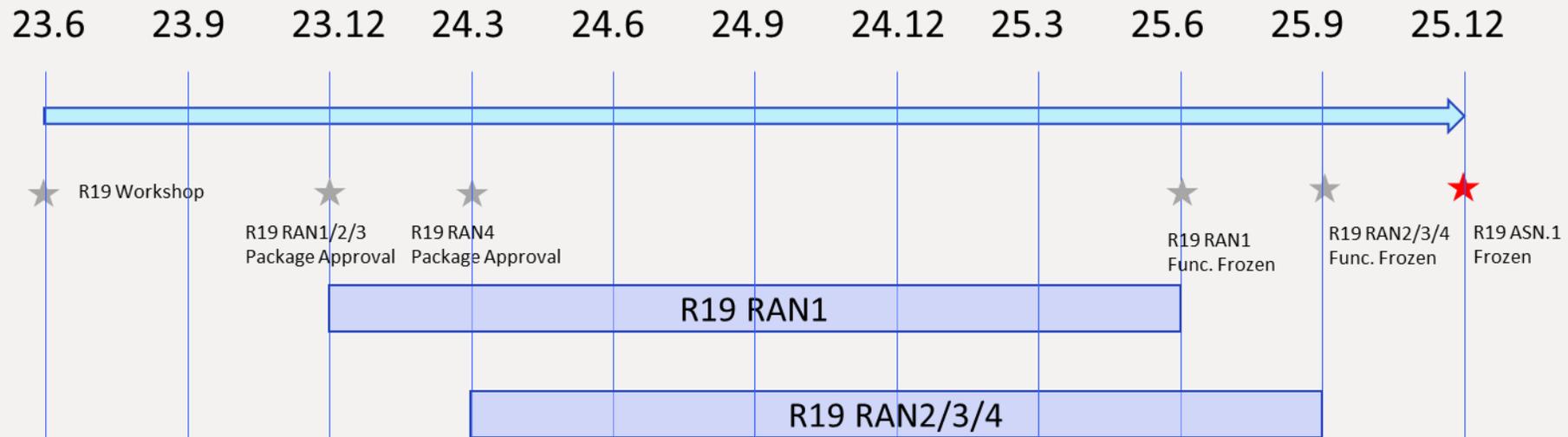
**Source: vivo**

**Title: Views on Rel-19 AI/ML for air interface**

**Document for: Discussion**

# Background: Expected RAN Rel-19 Timeline

Release timeline



- 18 months release duration is proposed for Rel-19 function completion
- Target Dec 25 for R19 ASN.1 frozen (3 months gap between func. Frozen and ASN.1 Frozen)

# Expected Areas for Rel-19 AI/ML

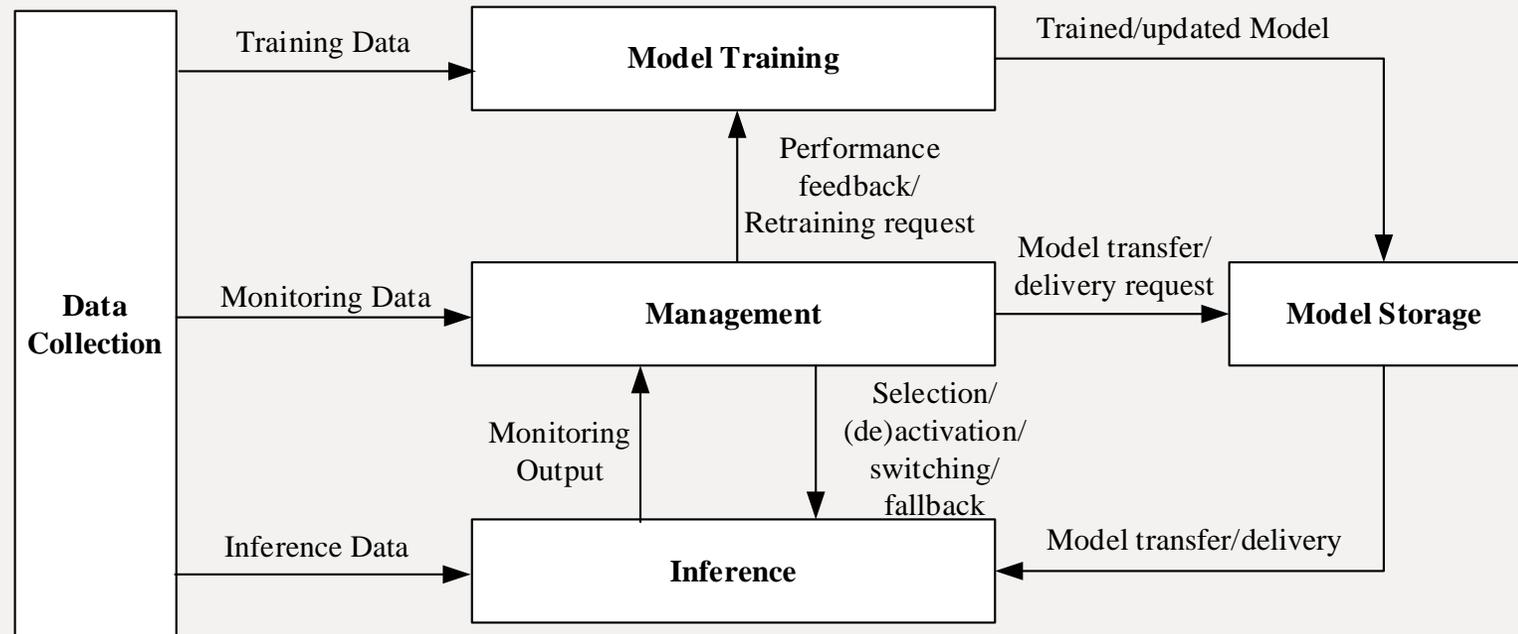
- Rel-18 continuation
  - Framework
  - Use case specific issues
- New use cases

# Rel-18 Continuation: framework

- In Rel-18 study, RAN1/RAN2/RAN4 identifies the following components for LCM and focus the study in part of the following:
  - **Data collection**
    - Note: This also includes associated assistance information, if applicable.
  - **Model training**
  - **Functionality/Model identification**
  - **Model transfer**
  - **Model inference operation**
  - **Functionality/Model selection, activation, deactivation, switching, and fallback operation.**
    - Including: Decision by the network (either network initiated or UE-initiated and requested to the network), decision by the UE (event-triggered as configured by the network, UE's decision reported to the network, or UE-autonomous either with UE decision reported to the network or without it)
  - **Functionality/Model monitoring**
  - **Model update**
    - Note: Terminology is to be defined. This includes model finetuning, retraining, and re-development via online/offline training.
  - **UE capability**

# Rel-18 Continuation: framework

- Based on the Rel-18 study on different LCM components, the following high level framework is expected to be outcome of the study:



# Rel-18 Continuation: framework

- Rel-19 AI/ML work for air interface is paving way for 6G, all the major LCM components well studied in Rel-18 should be included targeting a future-proof framework with the following assumptions.
  - Considering both the cases data collection is done at UE side and at the network side
    - For UE sided model, data collection is at UE side and/or network side;
    - For network sided model, data collection is at network side with assistance from UE;
    - For two sided model, data collection is at network side and/or UE side;
  - Considering both the cases model storage entities are at UE side or network side;
    - For UE sided model, model storage is at the network side and/or at UE side;
    - For network sided model, model storage is at the network side;
    - For two sided model, model storage is at the network side and/or UE side;
  - Considering both collaboration level y and level z
    - For the case that model is stored at the network, collaboration level z is envisioned;
- **Proposal: Consider a future-proof framework for Rel-18 continuation work in Rel-19 with the assumption that data collection can be done at both UE side and network side, model storage can be done at UE side and network side and both collaboration level y and level z are considered.**

# Rel-18 Continuation: framework

- For collaboration level z, different model transfer cases are identified.
  - Case z1/z2 are using proprietary format while Case z3/z4/z5 are using open format which is mutually recognizable between parties.
  - SA2 has already identified a token based format alignment solution between parties.
- All different model transfer cases can rely on same 3GPP solution(s)
  - E.g., both case z2 and case z4 are using the same CP solutions for CSI/beam, the same LPP based solutions for positioning;
- Reference model structure is being studied in RAN4
- Different model transfer cases (z1~z5) may have different assumptions on coordination needed between 3GPP entities and UE side;
- **Proposal: At least from RAN perspective, similar specification work can be done for different model transfer cases (z1~z5);**

# Rel-18 Continuation: Use cases

- The following use cases are studies in Rel-18:
  - CSI: CSI compression, CSI prediction.
  - Beam prediction: spatial domain, temporal domain;
  - Positioning: Assisted, Direct
- All use cases show gains compared with legacy;
  - CSI compression is with less gain compared to other cases but can be easily extended to time domain compression with much larger gains; (More results can be found in Appendix2)
  - CSI compression is the only two sided case among the six sub use cases;
- **Proposal: consider all six sub use cases studied in Rel-18 to be included in Rel-19 WI;**

# Rel-18 continuation: Relationship with SA

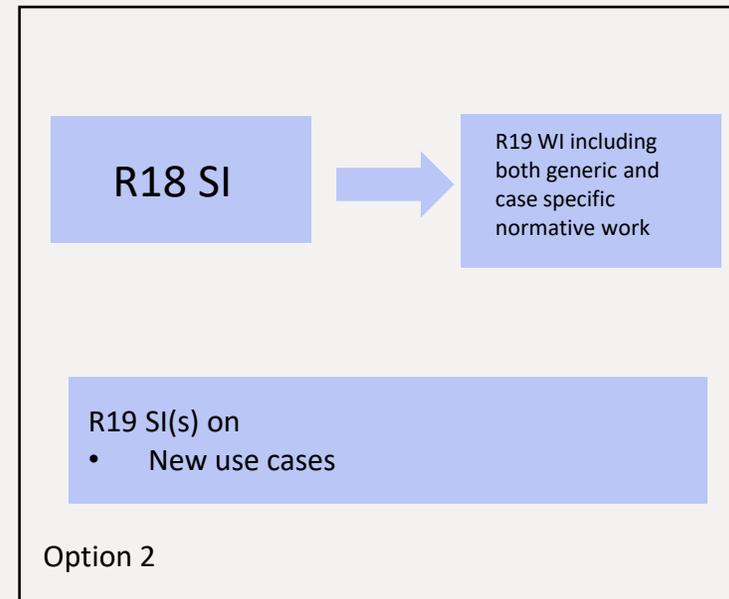
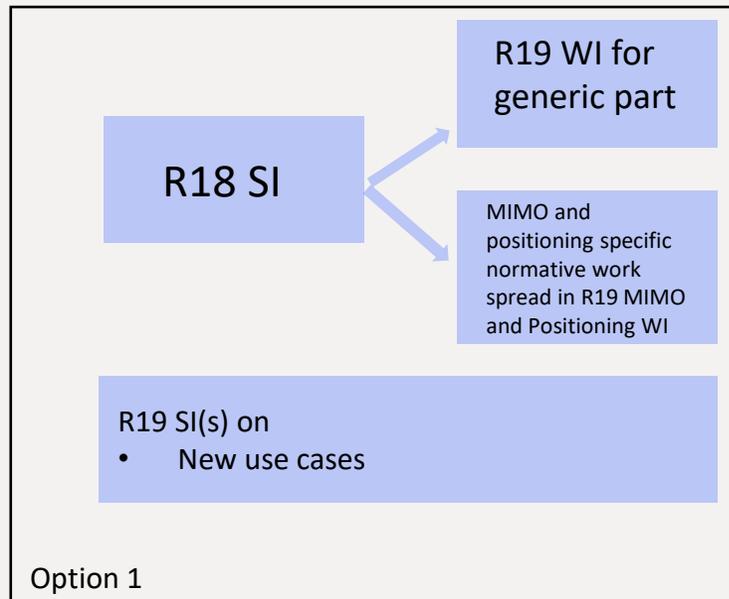
- At least the following areas are identified with SA impact for the three studied use cases:
  - AI + Positioning procedures and signaling
  - Model identification procedures and signaling
  - Model transfer/Data collection related issues;
- SA2 did not have discussion on related work in Rel-18:
  - SA2 Work expected would be large.
  - Rel-18 RAN study item did not trigger any SA2 work successfully. Lessons learned are that RAN should coordinate with SA as early as possible.
- **Proposal: RAN should coordinate with SA to study and specify the corresponding solutions for application of AI/ML in air interface.**

# New use cases for Rel-19 AI/ML for air interface

- Rel-18 studied PHY layer use cases. Rel-19 should consider more use cases that can further expand the application of AI/ML in air interface :
  - Higher layer use cases
  - Use cases that utilize features not directly related to channel
- Candidate use cases include:
  - **AI based mobility enhancements (RAN2-led) (More results can be found in Appendix3)**
  - **PA efficiency/nonlinearity improvement, including e.g., one-sided or two-sided operation (RAN1 or RAN4-led) (More results can be found in Appendix4 and joint contribution RWS-230240)**

# Arrangement of related work in Rel-19

- Rel-18 continuation work for the framework aspects is expected to be RAN2 leading
- Rel-18 continuation work for case specific aspects are correlated with specific procedure designed for MIMO and positioning, which would be RAN1 expertise.
- **Consider the following possible options on table for arrangement of new WI/SI in Rel-19.**



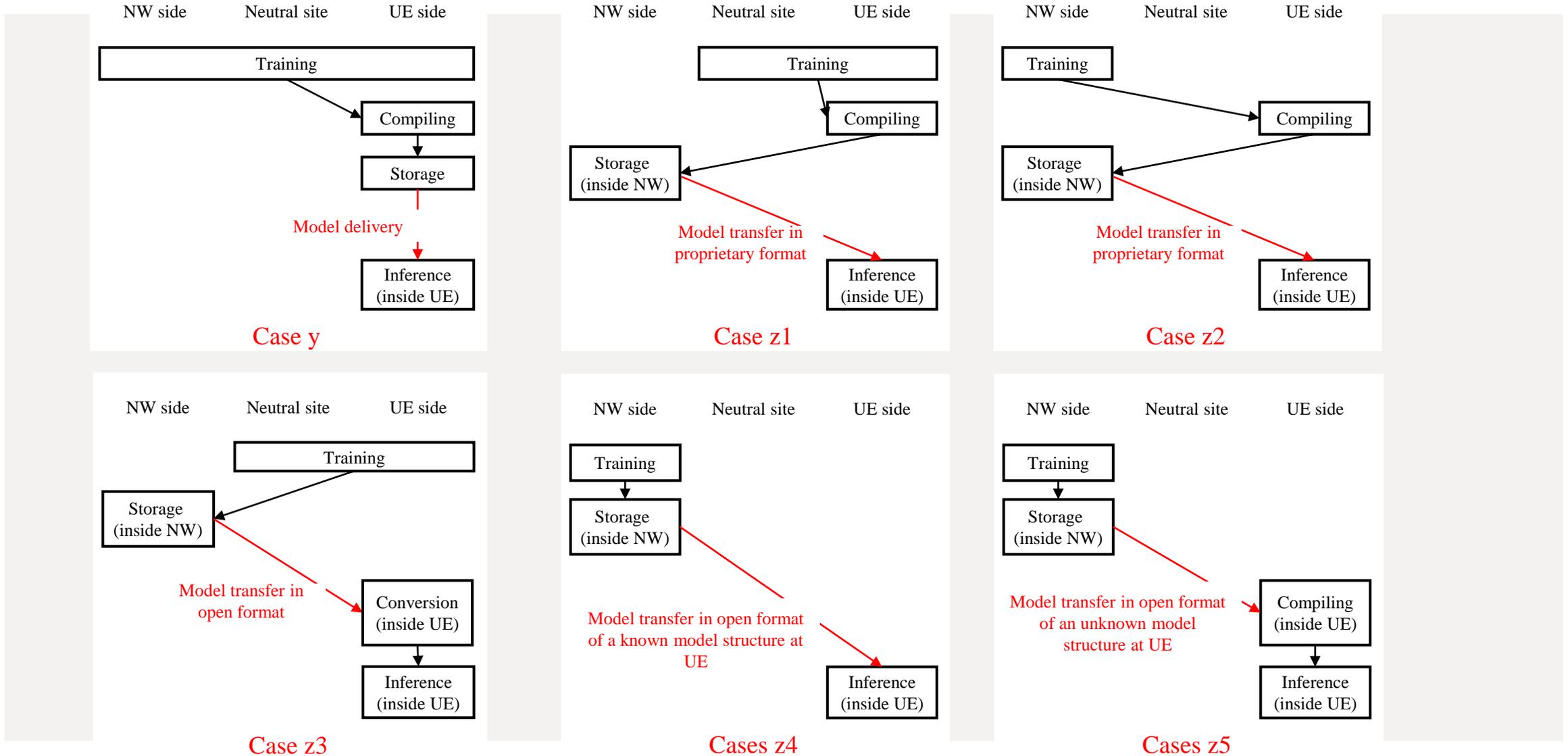
# Summary of views for Rel-19 AI/ML work

- **Consider a future-proof framework for Rel-18 continuation work in Rel-19 with the assumption that data collection can be done at both UE side and network side, model storage can be done at UE side and network side and both collaboration level y and level z are considered.**
- **At least from RAN perspective, similar specification work can be done for different model transfer cases (z1~z5);**
- **Consider all six sub use cases studied in Rel-18 to be included in Rel-19 WI;**
- **Consider another SI with more use cases to be included for a more comprehensive study:**
  - **Higher layer use cases**
  - **Use cases that utilize features not directly related to channel**
- **RAN should coordinate with SA to study and specify the corresponding solutions for application of AI/ML in air interface**

THANK YOU.

谢谢。

# Appendix1: Rel-18 discussion on model transfer cases



# Appendix1: Rel-18 discussion on model transfer options

⇒ **Agreed:** ↵

**Aim to at least analyze the feasibility and benefits of model/transfer solutions based on the following:**↵

**Solution 1a: gNB can transfer/deliver AI/ML model(s) to UE via RRC signalling.**↵

**Solution 2a: CN (except LMF) can transfer/deliver AI/ML model(s) to UE via NAS signalling.**↵

**Solution 3a: LMF can transfer/deliver AI/ML model(s) to UE via LPP signalling.**↵

**Solution 1b: gNB can transfer/deliver AI/ML model(s) to UE via UP data.**↵

**Solution 2b: CN (except LMF) can transfer/deliver AI/ML model(s) to UE via UP data.**↵

**Solution 3b: LMF can transfer/deliver AI/ML model(s) to UE via UP data.**↵

**Solution 4: Server (e.g. OAM, OTT) can transfer/delivery AI/ML model(s) to UE (e.g. transparent to 3GPP).**↵

## Appendix2: Temporal-spatial-frequency domain CSI compression

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

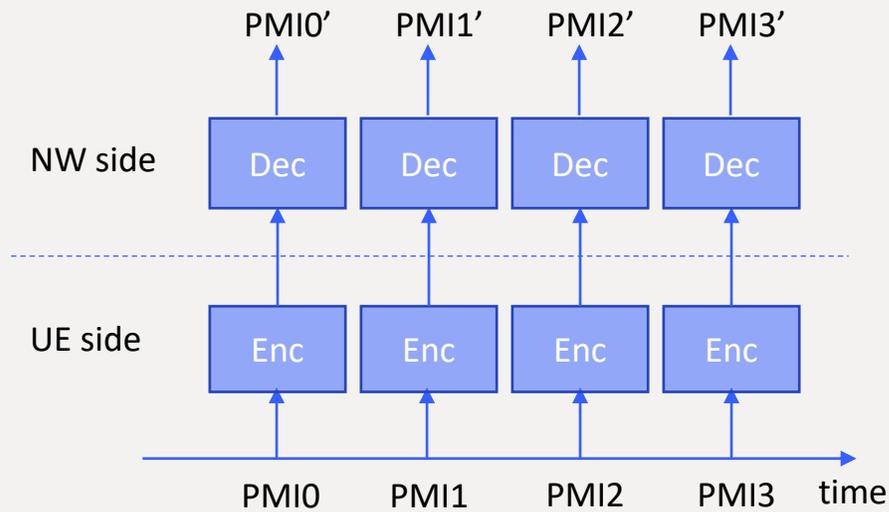


Illustration of spatial-frequency domain CSI compression in R18

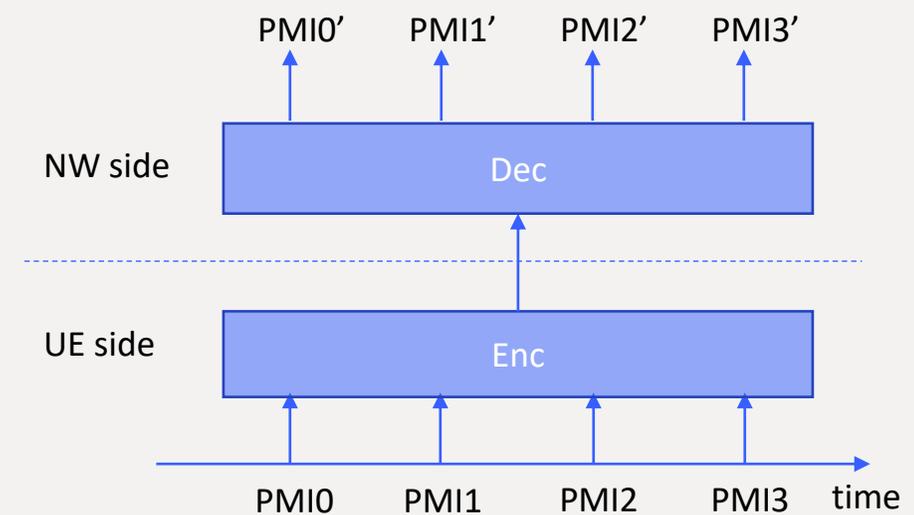


Illustration of temporal-spatial-frequency domain CSI compression

# Appendix2: Temporal-spatial-frequency domain CSI compression

- Initial results on joint compressing 4 PMIs (rank1 considered)

	SGCS on PMI0	SGCS on PMI1	SGCS on PMI2	SGCS on PMI3	Averaged SGCS
Legacy R16 codebook	0.8172	0.8170	0.8171	0.8172	0.8172
Benchmark: s-f model with payload 64*4	0.9071 (+11.02%)	0.9070 (+11.01%)	0.9074 (+11.05%)	0.9072 (+11.01%)	0.9072 (+11.01%)
Joint compression model with payload 64*4	0.9522 (+16.51%)	0.9610 (+17.62%)	0.9613 (+17.64%)	0.9573 (+17.14%)	0.9579 (+17.21%)
Joint compression model with payload 64*2	0.9250 (+13.21%)	0.9368 (+14.66%)	0.9393 (+14.95%)	0.9321 (+14.06%)	0.9351 (+14.42%)
Joint compression model with payload 64	0.8747 (+7.04%)	0.8933 (+9.33%)	0.8937 (+9.37%)	0.8796 (+7.63%)	0.8853 (+8.33%)

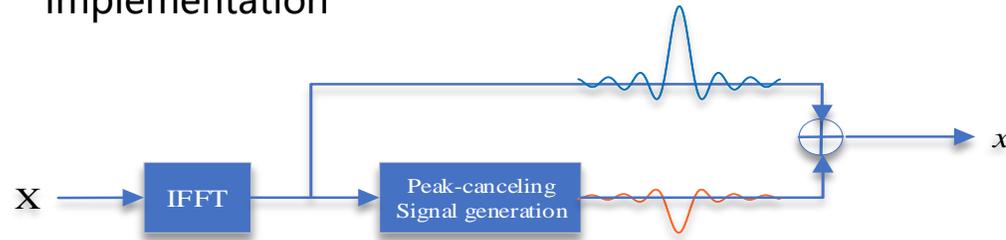
- Exploiting temporal domain correlation further provides ~6% relative SGCS gain on top of s-f domain compression, and the total improvement over legacy R16 CB is ~17% SGCS (~50% to ~60% additional gain)
- Overhead further reduces ~60% compared with s-f model, and the overall overhead reduction compared with R16 CB improves from ~30% to ~70%

# Appendix3: AI/ML for PA nonlinearity handling

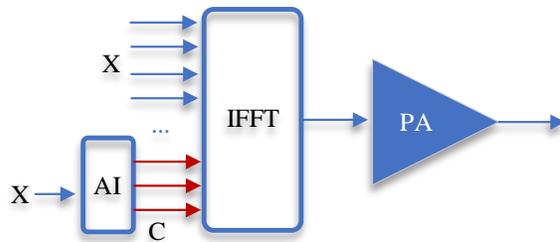
- Impact of PA non-linearity
  - Higher power back-off decreasing transmission power imposing UL coverage and throughput
    - In-band distortion: EVM
    - In-band emission: IBE
    - Out-of-band emission: ACLR, SEM
  - PA non-linearity impairment is heavier in large band width
- Popular methods to withstand PA non-linearity
  - PAPR reduction
    - TR: generate peak canceling signal for the reserved tones
    - SLM: select the scrambled signal from all scrambled signal set with lowest PAPR
    - etc.
  - Digital predistortion
    - Look up table: DPD based on AM In (V) – AM Out (V) shaping and AM (V) - PM Out (deg) shaping table
    - Volterra series: DPD formulated on Volterra series with estimated parameters
    - etc.

# Appendix3: AI + Tone Reservation for PAPR reduction

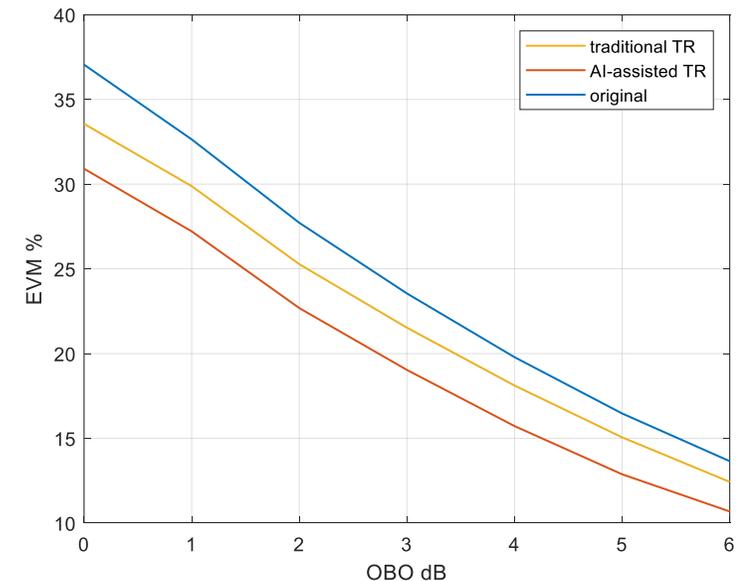
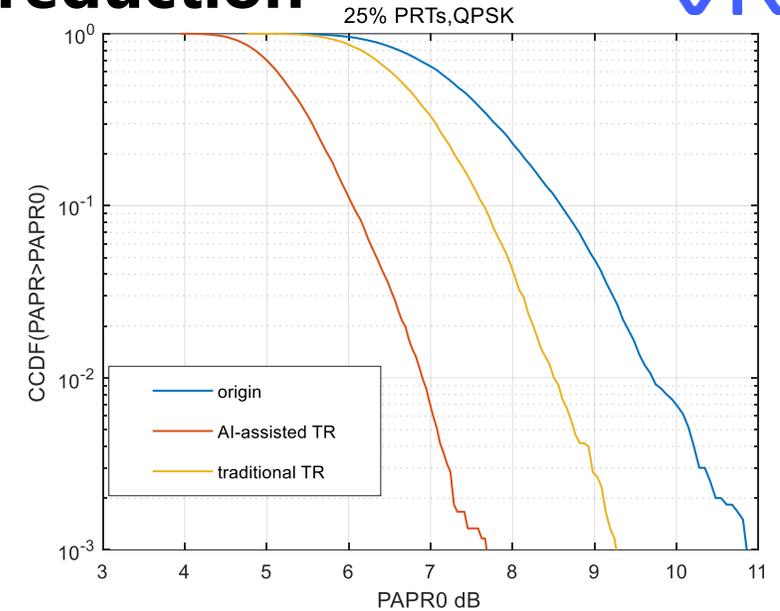
- TR (Tone reserve)
  - Implementation



- AI-assisted TR
  - AI-inferenced signal transmitting on reserved tones
  - AI-assisted peak canceling signal generation



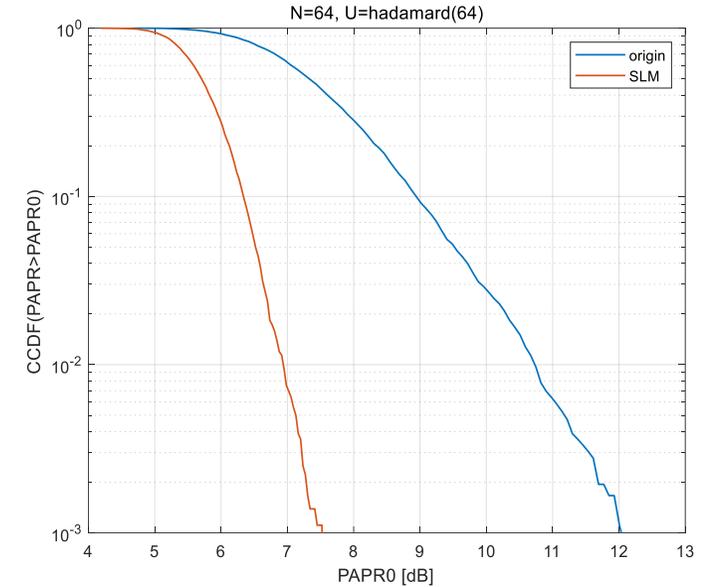
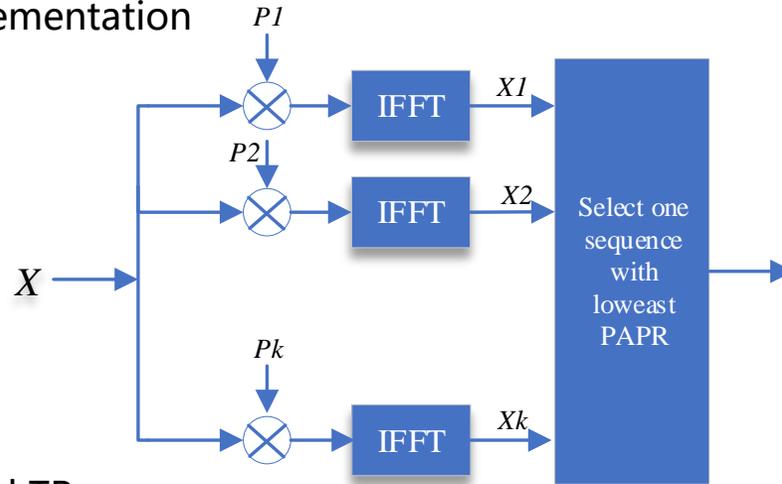
- Potential spec impact
  - Position and number of reserved tones should be informed to receiver



# Appendix3: AI + SLM for PAPR reduction

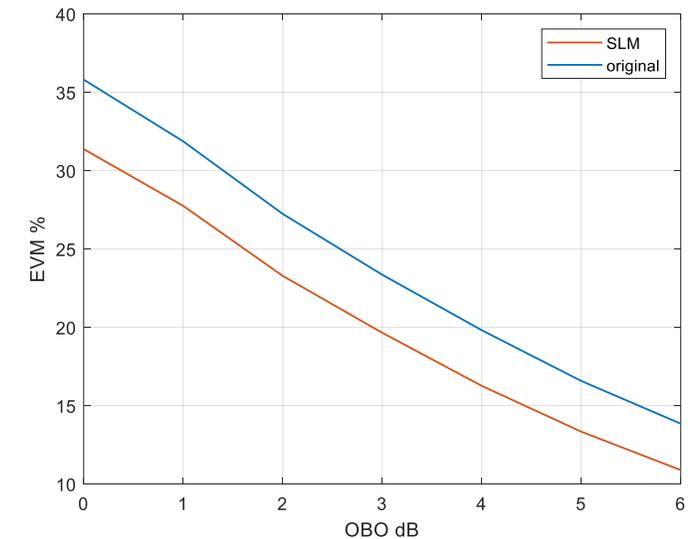
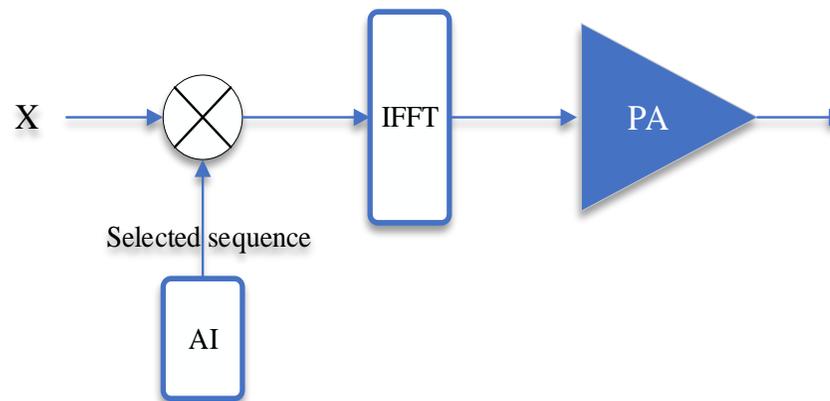
- SLM (selective mapping)

- Implementation



- AI-assisted TR

- AI-assisted sequence generation for selective mapping



- Potential spec impact

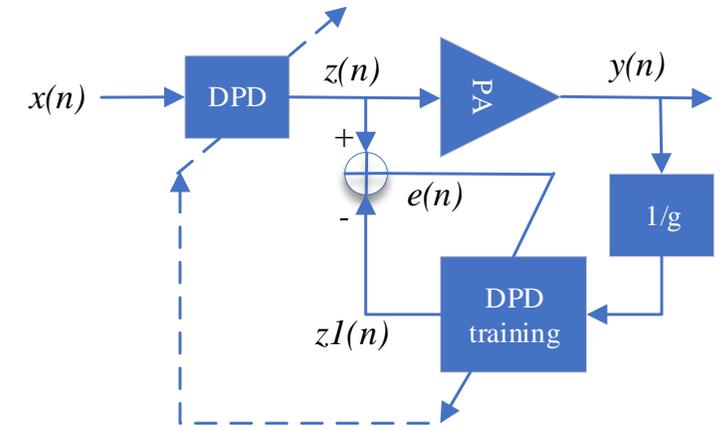
- Selected scrambling sequence should be informed to receiver for descrambling

# Appendix3: AI + DPD for PA nonlinearity optimization

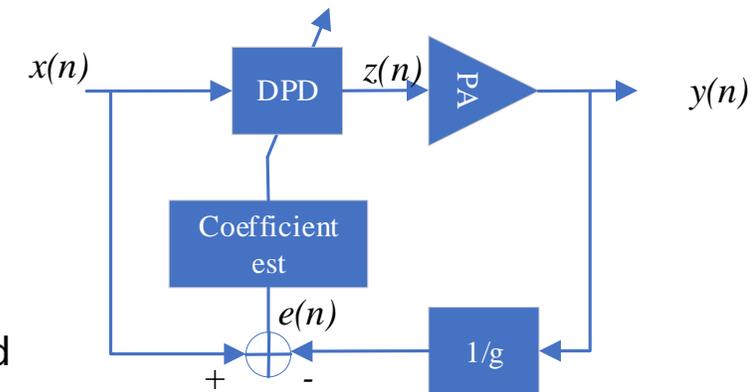
- DPD
  - Direct learning and indirect learning are considered
  - Performance evaluation

Schemes	NMSE	EVM	ACLR left	ACLR right
w/o DPD	4.7 dB	171%	-35.4 dB	-37.2 dB
Conventional DPD	-35.2 dB	1.73%	-41.2 dB	-42.3 dB
AI-based DPD	-46 dB	0.65%	-46.1 dB	-47.1 dB

- Potential spec impact is required to achieve adaptive DPD with feedback loop
  - Potential UL transmission delay due to DPD
  - Specific signal transmission for fine tuning with worse EVM/ACLR than required
  - etc.

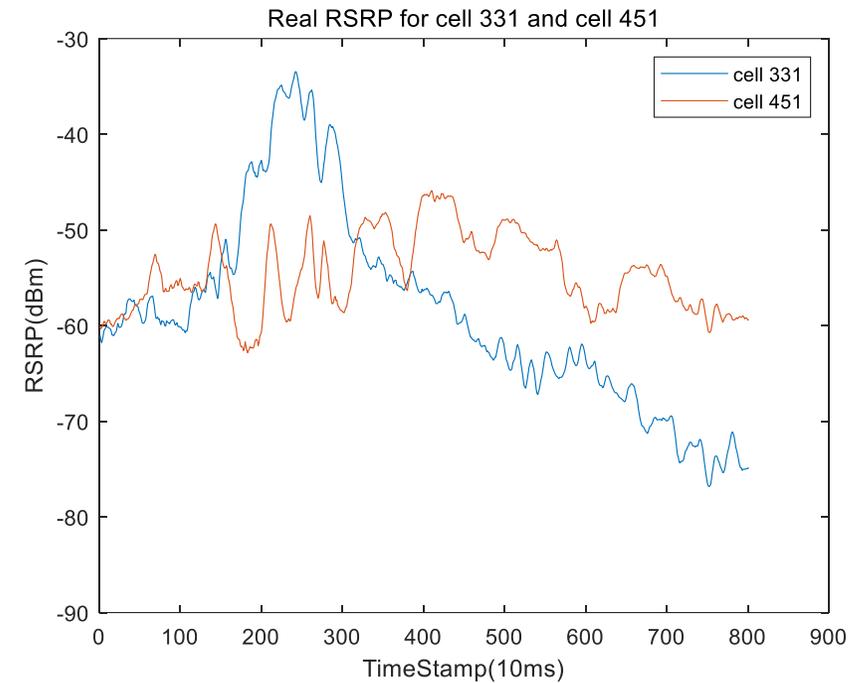


Indirect learning



Direct learning

# Appendix4: AI+Mobility field test

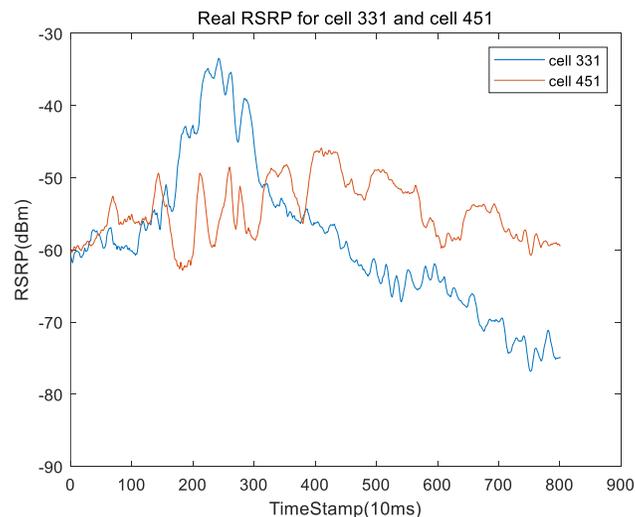


Carrier Frequency: FR1, 3.5GHz UE speed: 80~120km/h

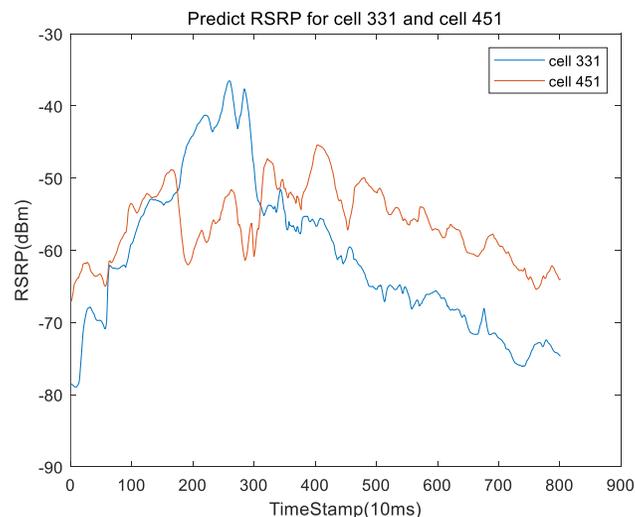
- During UE high-speed movement, the RSRP of the neighbor may becomes better than the serving cell in a short period of time.
- UE will handover to the neighbor cell and handover back quickly to the last serving cell, i.e., ping-pong handover occurs.

# Appendix4: AI+Mobility field test

## ■ Field test RSRP



## ■ Predict RSRP



## ■ UE speed: 80~120km/h

## ■ Dataset

- Training dataset: 15 UE trajectory
- Test dataset: 1 UE trajectory

## ■ AI Model

- Fully Connected Neural Network

## ■ Input

- History RSRP of serving & neighbor cell (interval = 10ms)
- History UE location and speed (interval = 1s)
- Observation window = 2s

## ■ Output

- Predict RSRP of serving cell
- Predict RSRP of neighbor cell
- Prediction timing = 2s

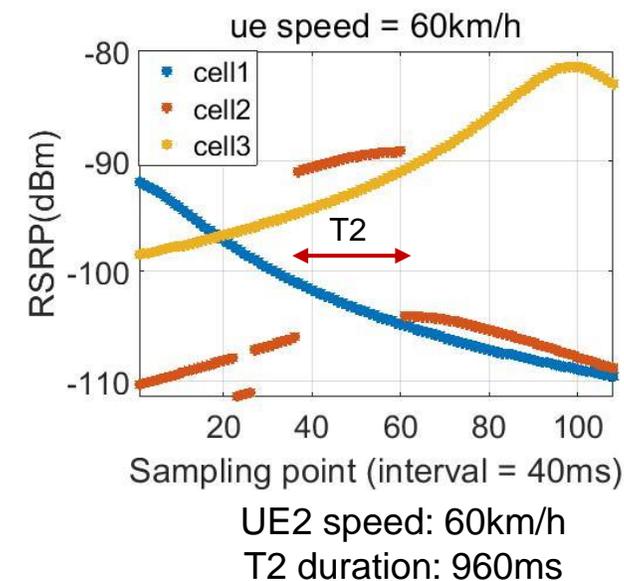
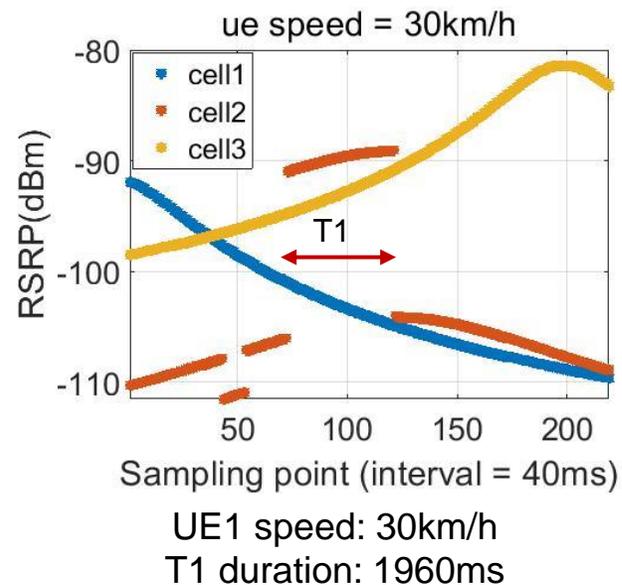
## ■ Performance

- RMSE = 0.8dBm

## Appendix4: AI+Mobility Ray tracing test



Simulation scenario and UE trajectory



- When the UE passes through the crossover, the RSRP of cell2 will change dramatically and UE handover to cell 2.
- For UE1 at a low speed, it will be served by cell 2 for a longer time (over 1 second).
- For UE2 at a high speed, it will only be served by cell 2 for less than 1 second.
- Both UEs may experience RLF when leaving the crossover and will reestablishment RRC connection on cell 3.

# Appendix4: AI+Mobility Ray tracing test

## ■ Accuracy of RRM measurement prediction

	Cell 1	Cell 2	Cell 3
Prediction 1	RMSE = 0.0044dB	RMSE = 1.08dB	RMSE = 0.26dB
Prediction 2	RMSE = 0.0062dB	RMSE = 1.11dB	RMSE = 0.26dB
Prediction 3	RMSE = 0.0844dB	RMSE = 1.23dB	RMSE = 0.28dB

Carrier Frequency: FR2, 30GHz

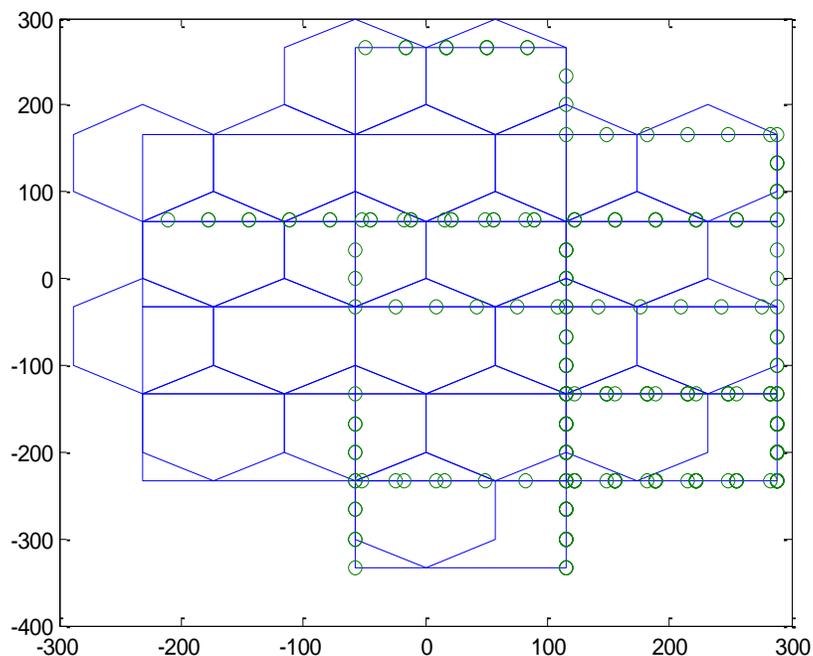
Prediction 1: RSRP of every 80ms in 320ms after  $T_0$

Prediction 2: RSRP of 1s after  $T_0$

Prediction 3: RSRP of 2s after  $T_0$

# Appendix4: AI+Mobility System evaluation test

## ■ Simulation scenario



Simulation scenario and UE trajectory

## ■ Simulation assumption

Attributes	Values or assumptions
Carrier Frequency	FR1: 4GHz; FR2: 30GHz
TRP Number	7 sites, 3 sector per site
Channel Model	3D-Uma in TR 38.901, support Spatial consistency ISD = 200m
UE speed	120km/h
Mobility management	Event: A3; Hysteresis: 2dB; Offset: 1dB; TimeToTrigger: 320ms, 40ms Handover preparation time: 50ms; Handover execution time: 40ms
RLM	L1 measurement period: 20ms Qin sliding window length: 100ms Qout sliding window length: 200ms Qin threshold: -6dB; Qout threshold: -8dB N310: 1; N311: 1; T310: 1s
Handover model and corresponding metrics	As defined in TR 36.839 Short time of stay: served by the target cell for less than 1s after HO

# Appendix4: AI+Mobility System evaluation test

## ■ RRM prediction based HO

		Legacy HO, TTT = 320	Legacy HO, TTT = 40	AI based HO	CHO, TTT = 320	CHO, TTT = 40	AI based CHO
FR1	HOF rate	9.16%	2.2%	1.95%	0.28%	0.15%	0.32%
	Ping-pong HO rate	1.1%	3.6%	0.37%	1.0%	3.7%	0.37%
	Short Time of Stay (1s) rate	13.4%	18.9%	5.7%	13.6%	18.8%	5.67%
FR2	HOF rate	7.4%	2.5%	2.0%	0.42%	0.43%	0.44%
	Ping-pong HO rate	5.2%	10.3%	2.7%	5.2%	10.3%	2.7%
	Short Time of Stay (1s) rate	24.1%	36.7%	10.4%	24.4%	36.5%	10.8%

- With RRM prediction, the unintended events rate during HO and CHO can be significantly reduced, including HOF rate, ping-pong HO rate and short time of stay rate.

# Appendix5: Rel-18 discussion on Model/functionality identification



Model identification for Type 1 training two-sided model and one-sided model transferred from network to UE

Model identification for Type 3 training two-sided model (separate training at two sides)

AI/ML-enabled Feature/FG + additional conditions for UE developed UE-sided model

Functionality identification

Step 0: Alignment of model structure. Network trains the model parameters based on collected data.

Step0: Alignment of model structure, quantization. Network/UE exchange data needed, trains the two-sided model and align on model/app ID.

Step0: UE collects the data needed for training and UE trains the model. Alignment of necessary information.

Step1: UE AI/ML-enabled Feature/FG report to network;

Step2a: Network delivers the UE-sided model parameters to UE, with other necessary information;

Step2b: Alignment of additional conditions (e.g., scenarios, sites, and datasets) between network and UE.

Step2b: Alignment of additional conditions (e.g., scenarios, sites, and datasets) between network and UE.

Step2b: Alignment of additional conditions (e.g., scenarios, sites, and datasets) between network and UE.

Step3: Using model ID or applicability ID, network controls LCM of the UE sided part of two sided models.

Step3: Using model ID or applicability ID, network can indicate or assist LCM, including model selection or switching.

Step3: Using applicability ID, network can indicate or assist LCM, including model selection or switching, model activation/deactivation.

Step2: Based on UE report and continuously monitored wireless conditions, network controls LCM of the AI/ML based functionality.

Step4: Additional conditions (e.g., scenarios, sites, and datasets) may be updated during usage.

Step4: Additional conditions (e.g., scenarios, sites, and datasets) may be updated during usage.

Step4: Additional conditions (e.g., scenarios, sites, and datasets) may be updated during usage.

