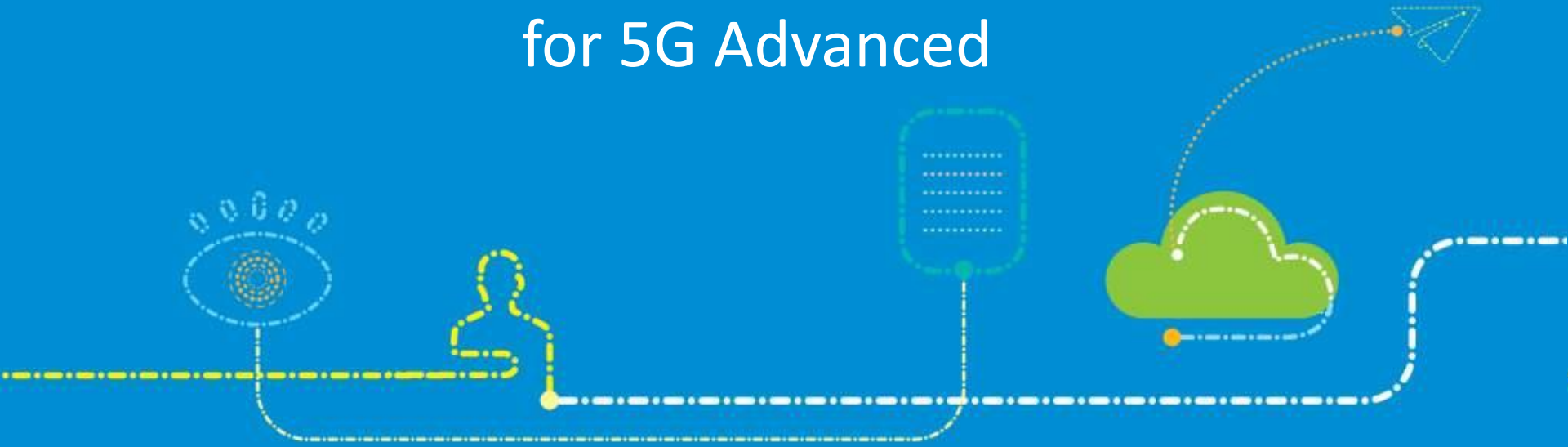
3GPP TSG RAN#94e
Electronic Meeting, Dec. 6 – 17

RP-213387

ZTE
Tomorrow never waits

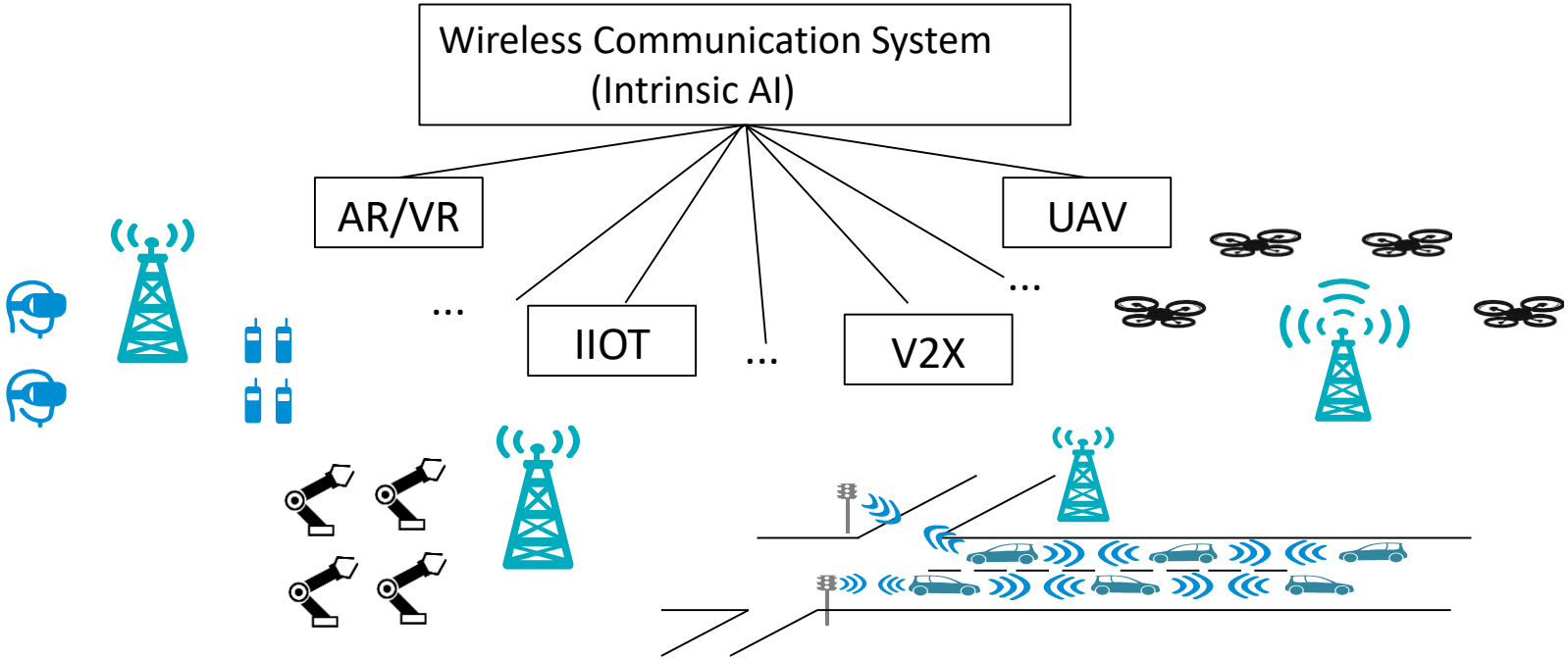
Source: ZTE, Sanechips
Agenda: 8A.1

Support of Artificial Intelligence Applications for 5G Advanced



Intrinsic AI

AI is expected to be the core functionality with standard support in 5G advanced networks



- Support AI everywhere - covering massive variety of use cases and all communication layers
  - AI for 5G advanced (Designing AI to enhance communication systems)
  - 5G advanced for AI (Designing communication systems for better support of AI based service)

Potential Benefits of AI standardization in PHY aspects

- AI algorithms are used in the current networks
  - However, limited applications in PHY due to complexity and lack of coordination between gNB and UE
- Enhancements (e.g. by providing assisted information) can be beneficial to achieve faster/more accurate learning.
- gNB and UE can collaborate on training and execution to reduce complexity.
- Machine learning used in physical layer has been drawn a lot of interest in recent research.
- The potential benefits are, e.g. :
  - Acquiring more accurate channel information with less CSI feedback overhead
  - RS overhead can be reduced for CSI acquisition/beam management
  - Reduce control signalling overhead
  - More accuracy for positioning
- These potential benefits translate to better system performance e.g. in terms of throughput and reliability

© ZTE Corporation. All rights reserved

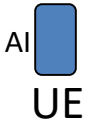
Use case and AI model characterization

- **CSI feedback enhancement:** Study beam prediction in time or spatial domain to improve UE experience in FR2 with reduced the beam measurement RS overhead and UE measurement/reporting.
  - For beam prediction in time domain, focus on high mobility scenario (e.g., high-speed train and high-way)
  - Study aspects include evaluate the feasibility and potential system level gain and specification enhancements such as beam measurement, beam report, beam indication and fallback approach
- **Beam management:** Study channel matrix or precoding matrix compression feedback in at least spatial domain and frequency domain
  - Use Rel-16 eType II codebook as the baseline scheme
- **Positioning accuracy enhancements:** Study positioning accuracy enhancement in scenarios with high NLOS probability, e.g., IIOT DH.
- More details on use case study can be found in Appendix
- **AI model management** needs to be studied considering the use cases and the collaboration levels in P5.
  - Model management includes the whole model life cycle, e.g., model deployment, training, update, etc.
  - Strive for a unified model management framework

© ZTE Corporation. All rights reserved

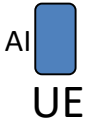
Collaboration level categorization

AI model is deployed in gNB or UE side through implementation without collaboration.



CAT 0a

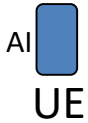
gNB provides information/reference signal to UE to assist UE AI operation, or the other way around. AI model parameters are not impacted by such information, e.g., the information just impacts AI input.



CAT 0b

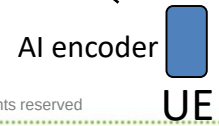
Note: These collaboration levels may involve both interference and training

gNB provides information to UE to online update UE AI model, or the other way around. AI model parameters are updated online by such information.



CAT 1

UE-gNB collaborative AI operation for split inference. AI model alignment is needed.



CAT 2

© ZTE Corporation. All rights reserved

Evaluation methodology

- The general principle of evaluation methodology shall follow 3GPP general framework, including system-level or link-level based on the target use case.
- Data set is generated based on 3GPP channel model in TR38.901, but parameter randomization shall be taken into account in training and inference.
    - System-level channel model is recommended for both SLS and LLS.
    - For system-channel model, parameter randomization is still needed for training and inference including some large-scale parameters, like delay spread, angular spread, noise power, etc., so that independent parameters can be used for training and inference.
- Assumption for collaboration level and training procedure should be discussed or reported by companies.
    - Collaboration levels in P5 can be considered.
    - Training procedure can be online training or offline training. Both can be evaluated considering performance, air interface overhead, UE and gNB complexity, etc.
- Used AI model also needs to be reported by companies.
    - One or a small set of reference models can be aligned among companies at least for calibration.

© ZTE Corporation. All rights reserved

Comments on draft SID RP-212708

Use cases to focus on:

•Initial set of use cases includes:

-CSI feedback enhancement, e.g., overhead reduction, improved accuracy, prediction [RAN1]

-Beam management, e.g., beam prediction in time, and/or spatial domain for overhead and latency reduction, beam selection accuracy improvement [RAN1]

-Positioning accuracy enhancements for different scenarios including, e.g., those with heavy NLOS conditions [RAN1]

~~-RS overhead reduction [RAN1]~~

~~-RRM Mobility, e.g., prediction in time or frequency for robustness, interruption and overhead reduction [RAN2]~~

•Finalize ~~representative~~ initial set of use cases ~~(reduced from the initial set and minimizing sub use cases)~~ for characterization and baseline performance evaluations

•Evaluate performance benefits of AI/ML based algorithms:

-Methodology based on statistical models (from TR 38.901 and TR 38.857 [positioning]), for link and system level simulations.

○Extensions of 3GPP evaluation methodology for better suitability to AI/ML based techniques should be considered as needed.

~~○Whether f~~Field data may be optionally used ~~are needed~~ to further assess the performance and robustness in real-world ~~environments should be discussed as part of the study~~.

○User data privacy needs to be preserved.

○Need for common dataset construction for training, validation and test for the selected use cases

○Consider adequate model training strategy, collaboration levels and associated implications, e.g., offline training vs. online training of models

•Assess potential specification impact, specifically for the agreed use cases in the final representative set and for a common framework:

-…

-Protocol aspects including (~~Except use case study,~~ RAN2 only start following general assessment after there is sufficient progress on use study in RAN1)

○ Consider aspects related to, e.g., capability indication, configuration procedures (training/inference), validation and testing procedures, and management of data and AI/ML model per RAN1 input

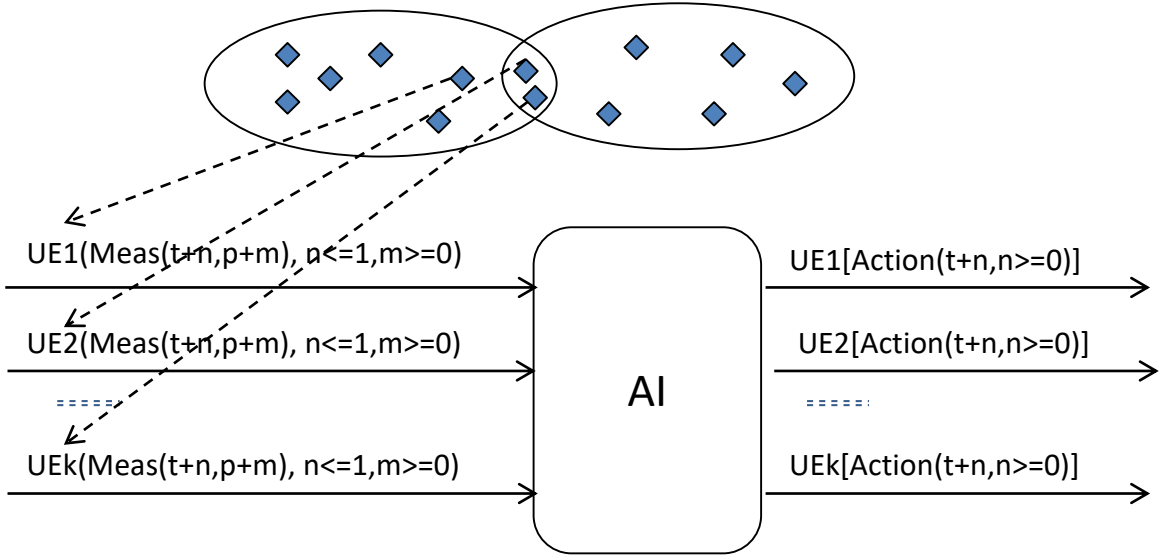
○Collaboration level specific specification impact per use case including signalling design to support the collaboration identified in RAN1

© ZTE Corporation. All rights reserved

Appendix: details on use case study

© ZTE Corporation. All rights reserved

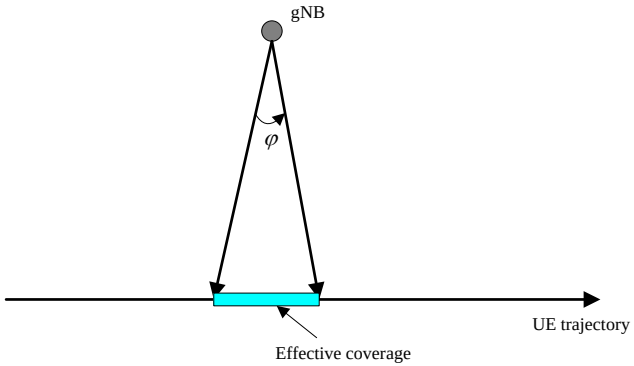
AI for Mobility and Beam Management



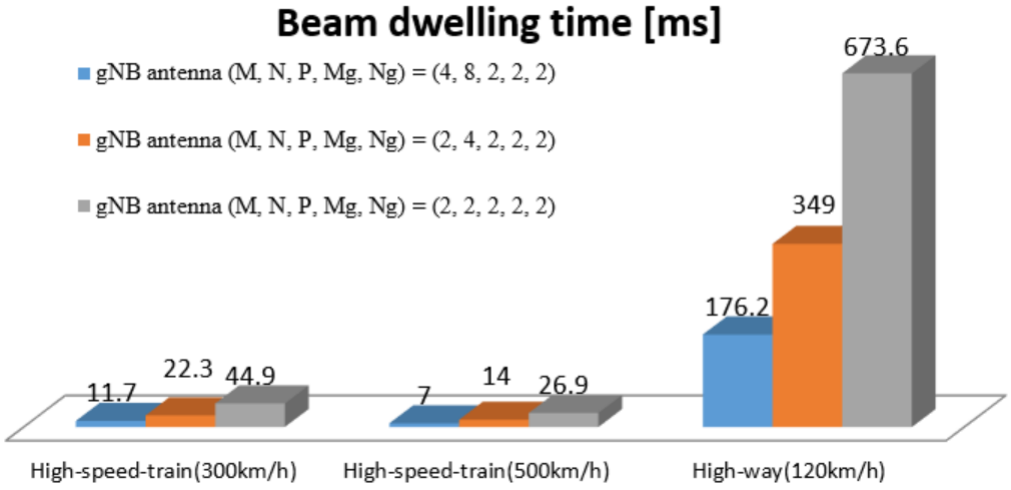
- AI model: Action for now(t0) or prediction for future(t1,t2...) for one UE is deduced based on measurement/event from related UEs jointly by taking their history(t-1,t-2...) and position stamp (p+m) into account. e.g. scheduling/beam tracking at gNB
- UE may predict future event/measurement (t1) by running AI algorithm based on measurements in the past (t-1,t-2). Predicted result is fed back to gNB e.g. mobility/beam prediction at UE

© ZTE Corporation. All rights reserved

Predictable Mobility for Beam Management



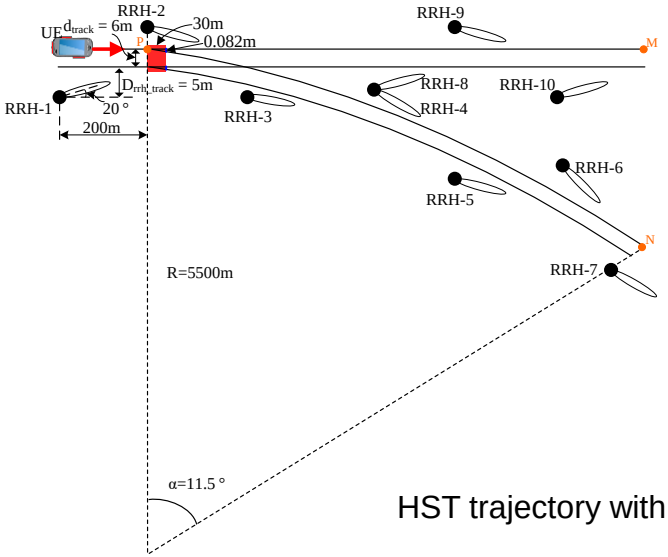
Effective coverage depending on beam-width, distance and speed



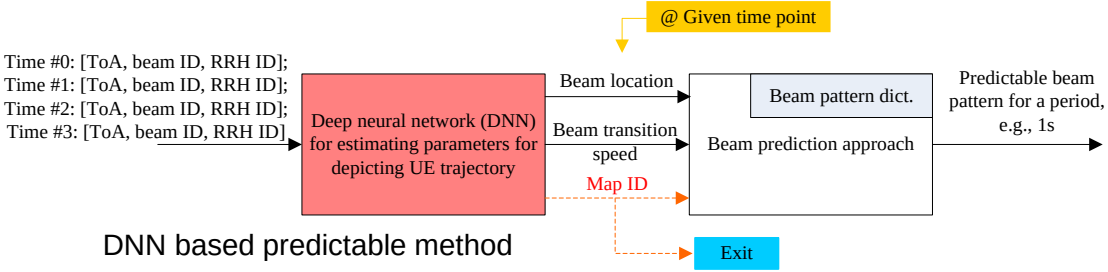
- Beam dwelling time depends on multiple factors including UE speed, distance between gNB and UE and beam-width of the beams.
- It can be observed that the beam dwelling time can be as small as 7ms. With the current beam management procedure including beam reporting, beam group activation and beam indication, it is hard to meet this beam update requirement. To make sure narrow beams can be used for better coverage and performance in high speed scenarios, it is beneficial to use AI on beam prediction together with trajectory prediction for mobility.
- Potential standardization impacts include enhancements on beam reporting, beam activation and beam indication.

© ZTE Corporation. All rights reserved

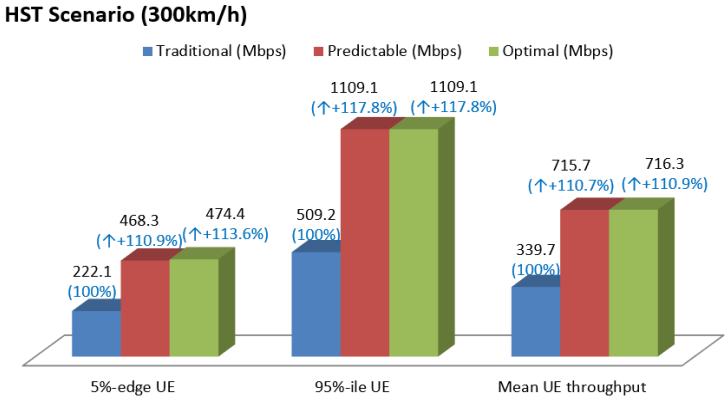
Predictable Mobility for Beam Management



HST trajectory with RRHs



DNN based predictable method



- When mobility is predictable, beam management can be also predictable.
- With predictable beam management based approach, it can be observed that significant performance gain is obtained due to RS overhead reduction and/or narrow beams.
- Details can be found in RWS-210481.

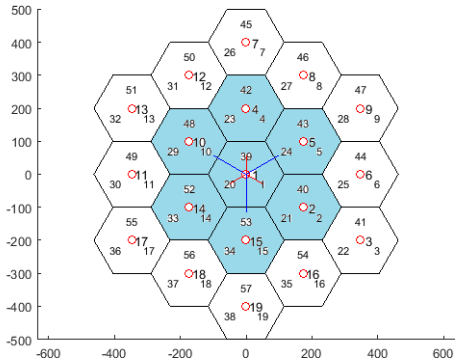
Table 1 Successful beam detection ratio compared with optimal beam

Scenario	Traditional	Predictable (non-AI)	Predictable (AI)
Line	93.46%	93.6%	99.2%
Curve	93.84%	94.37%	96.9%
Total	93.65%	93.99%	98.06%

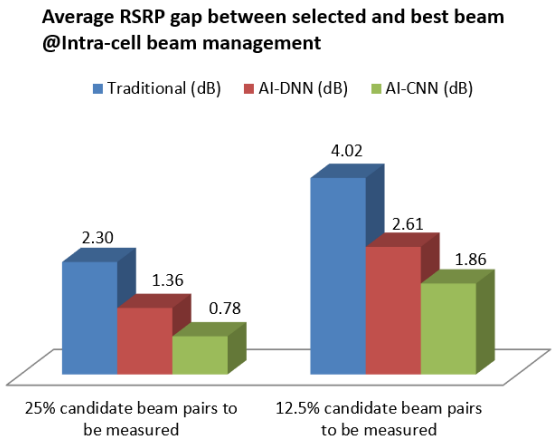
© ZTE Corporation. All rights reserved

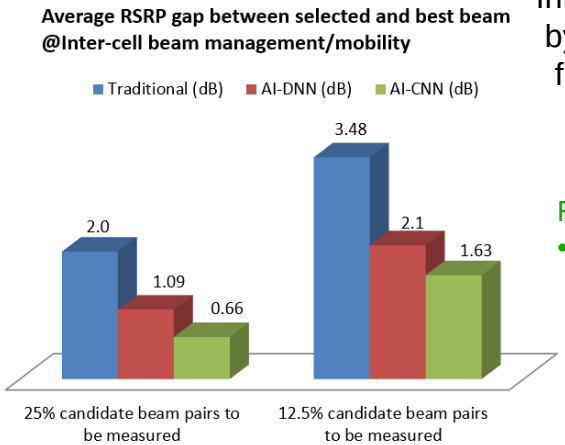
Compressed beam training for intra/inter-cell Beam management/mobility

- Measure a part of candidate beam pairs, and use AI to estimate qualities for all candidate beam pairs. Based on the estimate, the ‘best’ beam pair (as well as cell index) is selected for subsequent transmission
- The same AI-CNN/DNN can be used for both intra-cell/inter-cell scenarios (i.e., CNN/DNN trained in intra-cell scenario can be directly used for inter-cell directly with good generalization capability)
- It is observed that compressed with traditional method, AI DNN/CNN can significantly reduce the average RSRP gap between selected and best beam pair, especially for inter-cell.



Inter-cell beam management/mobility by measuring candidate beam pairs from serving and neighboring cells



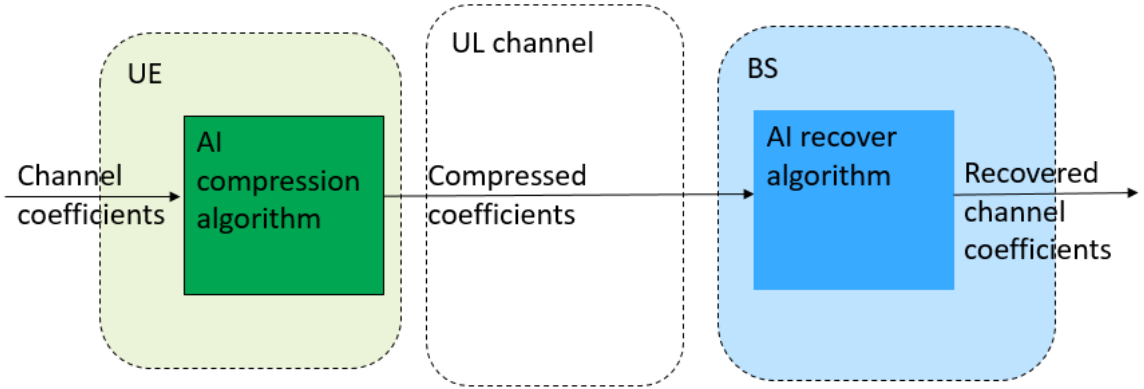


Remark:

- For traditional scheme, the best beam pair is selected from measured beam pair, e.g., identify best beam from 25%/12.5% candidate beam pairs.

CSI compression feedback based on AI: general procedure

- AI algorithm can compress the channel coefficient vectors in both spatial and frequency domain to a new coefficient vector with smaller size, to optimize the system performance v.s. overhead.
  - More accurate CSI feedback can enable better MU operation for massive MIMO
- For a trained AI algorithm in UE or BS side
  - At UE: Input is the original channel coefficients measured from P CSI-RS ports in M RBs, and output is a new coefficient vector with smaller size. The size of the new coefficient vector depends on compression ratio, which is the ratio between the overhead to report the original channel matrix and the overhead to report the compressed coefficients.
  - At BS: Input is the compressed coefficient vector received from UE, and output is the recovered channel coefficients for all the M RBs.



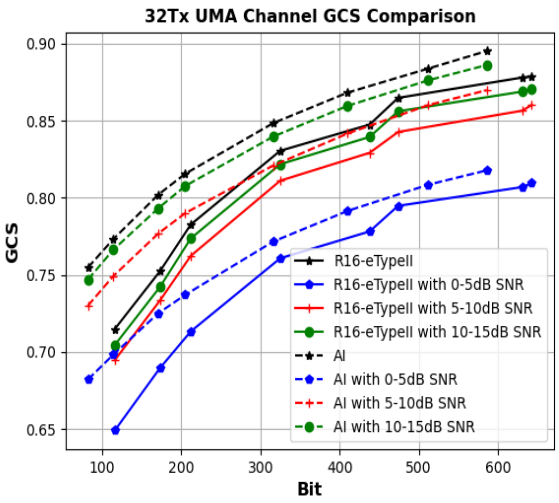
© ZTE Corporation. All rights reserved

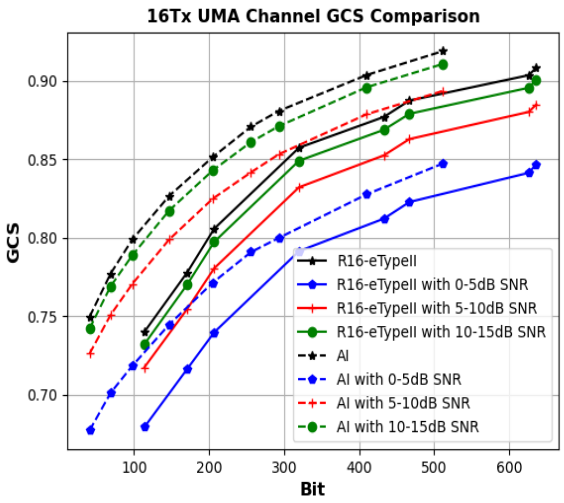
CSI compression feedback based on AI: evaluation set-up

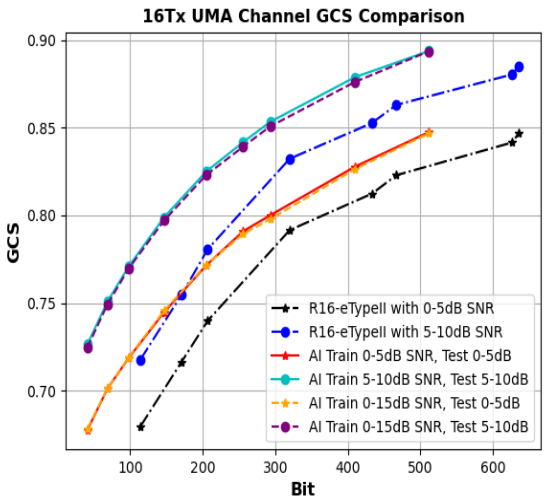
- Evaluations are conducted to show the benefit of AI based approach to compress and feed back CSI.
- BS has 16/32 Tx antennas. UE has one Rx antenna. The bandwidth has 128 RBs with 4-RB subband size.
- Channel model follows 38.901 system–level model. Noise power is random within an SNR range.
- Baseline is the Rel-16 eType II codebook to compress channel matrix in spatial and frequency domain.
- X axis is the number of feedback bits. Y axis is Generalized Cosine Similarity (GCS) between real channel $\mathbf{v}_1$ and gNB recovered channel $\mathbf{v}_2$ where $\mathrm{GCS}(\mathbf{v}_1, \mathbf{v}_2) = \frac{|\mathbf{v}_1 \mathbf{v}_2^{\mathrm{H}}|}{\|\mathbf{v}_1\|_2 \cdot \|\mathbf{v}_2\|_2}$.
- It can be observed from simulation results in the next two pages that, for a given performance level, AI based approach shows significant overhead saving compared with Rel-16 eType II. Similarly, for same number of feedback bits, large performance gain can be achieved. These reveal the GREAT potential of AI based approach.

© ZTE Corporation. All rights reserved

CSI compression feedback based on AI: Evaluation result-1





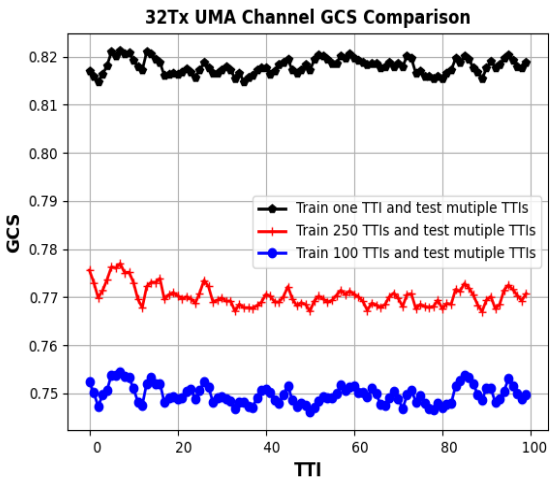


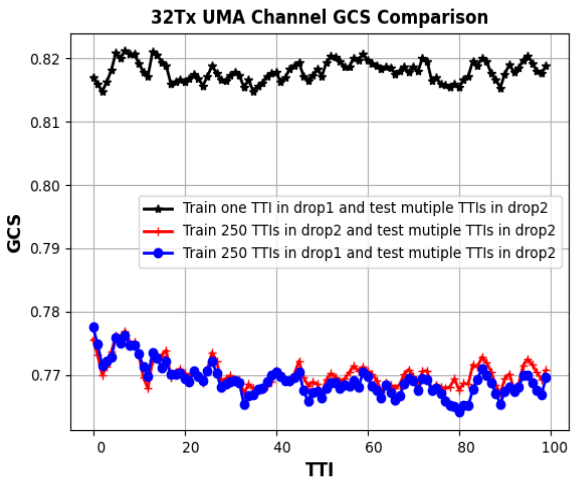
Observation:

- AI model shows good performance even there is noise in channel estimation.
- AI model shows good robustness even the SNR of training and SNR of inference are different.

© ZTE Corporation. All rights reserved

CSI compression feedback based on AI: Evaluation result-2





Remark:

- Large-scale parameters (e.g., angular spread and delay spread) are generated independently for different drops, subject to a same distribution.
- Each black curve trains 90000 geometric points, while the blue and red curves train 420 geometric points

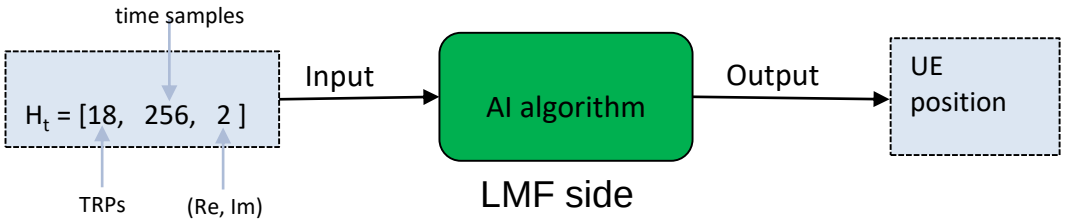
Observation:

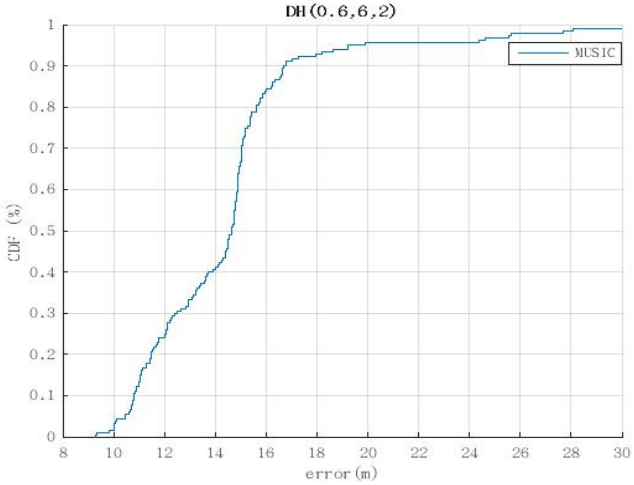
- AI model shows robust performance against channel time variation.
- Offline training can provide good performance, when the training of geometric points is sufficient.
- AI model can provide good robustness even when some large-scale parameters are not same for training and inference.

© ZTE Corporation. All rights reserved

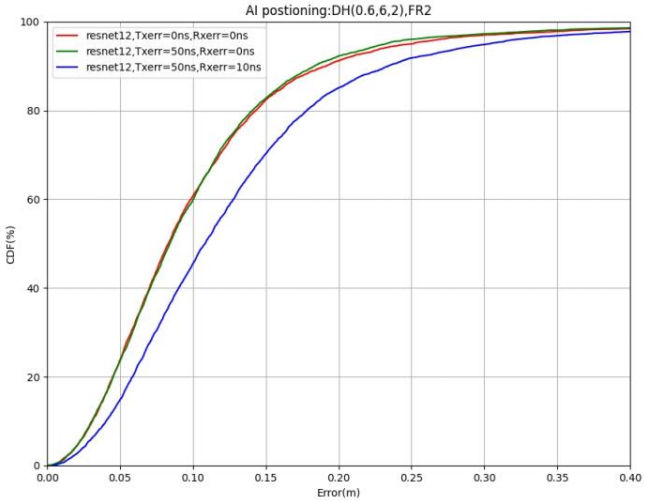
Positioning based on AI

- AI can significantly improve positioning accuracy in scenario with low LoS probability

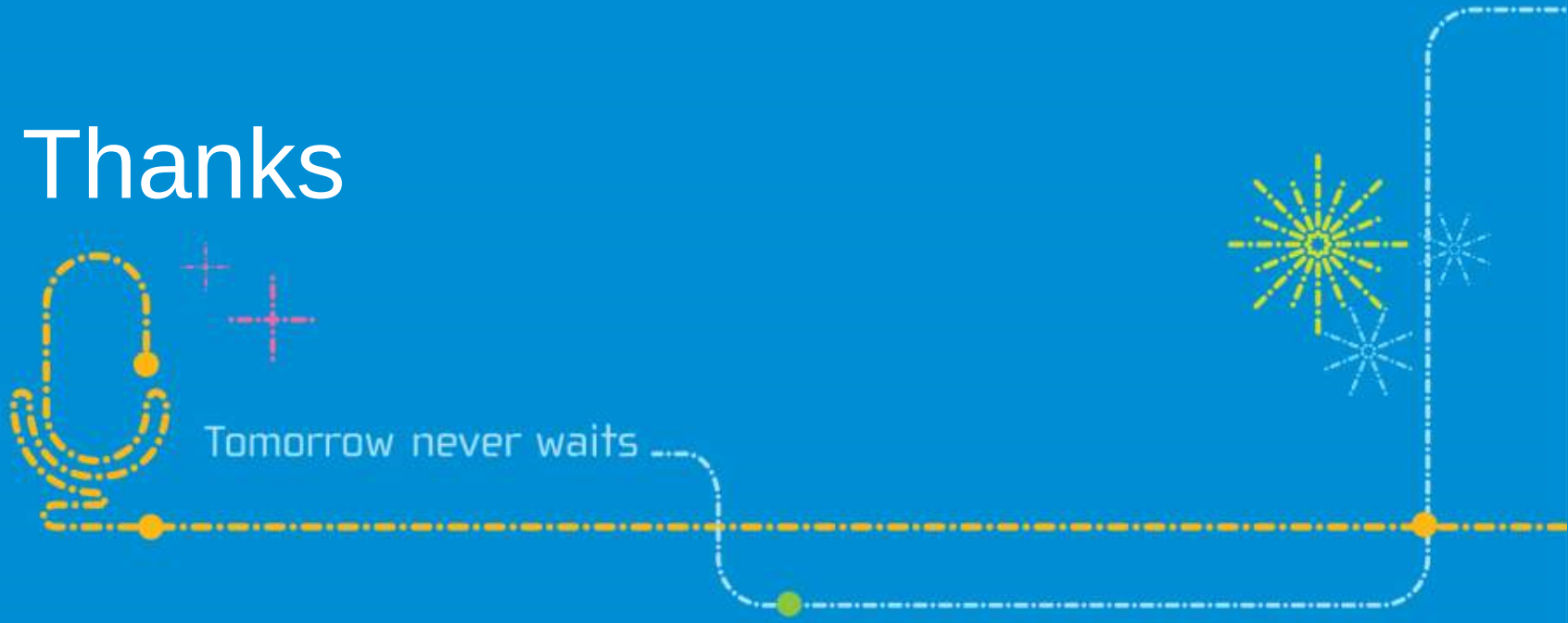




vs.



- 15m positioning errors by traditional positioning solution in DH scenario with very high NLOS probability
- 0.2m positioning errors by AI positioning in DH scenarios with very high NLOS probability and/or with 50ns synchronization errors

© ZTE Corporation. All rights reserved

Thanks