**3GPP TSG RAN WG1 #118bis R1-2408952**

Hefei, China, October 14th – 18th, 2024

**Agenda Item: 9.1.4.1**

**Source: IIT KANPUR**

**Title: Discussion on additional study of AI/ML for CSI compression**

**Document for: Discussion and Decision**

# Introduction

In RAN#102, the new WID for AI/ML for the NR air interface was finalized. Further, in RAN#105 meeting, the WID for Rel-19 AI/ML for the NR air interface was updated where, CSI prediction will be specified in Rel-19, while further study on CSI compression will continue through the end of Rel-19.

The WID outlines additional study on CSI compression, with the following key objectives [1]:

|  |
| --- |
| Study objectives   * CSI feedback enhancement [RAN1]:   + For CSI compression (two-sided model), further study ways to:     - Improve trade-off between performance and complexity/overhead       * e.g., considering extending the spatial/frequency compression to spatial/temporal/frequency compression, cell/site specific models, CSI compression plus prediction (compared to Rel-18 non-AI/ML based approach), etc.     - Alleviate/resolve issues related to inter-vendor training collaboration.   while addressing necessary specification impact analysis, as well as, other aspects requiring further study/conclusion as captured in the conclusions section of the TR 38.843. |

Up to RAN1#115, several agreements were reached regarding the study of AI/ML for CSI compression, and the detailed analysis of this study is documented in TR 38.843 [2]. Furthermore, by RAN1#116bis, additional agreements concerning evaluation assumptions and inter-vendor collaborations were established [3].

In this contribution, we present our observations based on evaluation results related to T-S-F domain compression for AI/ML-based CSI compression and our views on inter-vendor collaboration.

# Temporal-spatial-frequency domain compression evaluations

In this section, we describe details about the T-S-F domain compression evaluations and our initial results and observations.

The relevant agreements regarding evaluation assumptions in RAN1#116 and RAN1#116bis are shown below:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Agreement**  For the evaluation of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Release 19, adopt the following categorization for study:   |  |  |  |  | | --- | --- | --- | --- | | Case | Target CSI slot(s) | Whether the UE uses past CSI information | Whether the network uses past CSI information | | 0 | Present slot | No | No | | 1 | Present slot | Yes | No | | 2 | Present slot | Yes | Yes | | 3 | Future slot(s) | Yes | No | | 4 | Future slot(s) | Yes | Yes | | 5 | Present slot | No | Yes |   **Agreement**  For the evaluation of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Release 19, adopt the following as baseline options for UE distribution:   * Option 1: 80% indoor, 20% outdoor * Option 2: 100% outdoor   Note: Indoor speed is 3 km/h, outdoor speed is chosen from the following options: 10 km/h, 20 km/h, 30 km/h, 60 km/h, 120 km/h. Assumption on O2I car penetration loss and spatial consistency follow the R18 AI based CSI prediction.  **Agreement**  For the evaluation of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Release 19, adopt the following evaluation assumptions:   * CSI-RS configuration   + Periodic: 5 ms periodicity (baseline), 20 ms periodicity(encouraged)   + Aperiodic (for cases with prediction): Optional, CSI-RS burst with K resources and time interval m milliseconds (based on R18 MIMO eType-II) * CSI reporting periodicity: {5, 10, 20} ms; other values are not precluded * For cases with the use of past CSI information, to report observation window, including number/time distance of historic CSI/channel measurements. * For cases with prediction, to report prediction window, including number/time distance of predicted CSI/channel.   **Agreement**  For the evaluation of AI/ML-based CSI compression using localized models in Release 19, consider the following options as a starting point to model the spatial correlation in the dataset for a local region:   * Option 1: The dataset is derived from UEs dropped within the local region, with spatial consistency modelling as per TR 38.901.   + - E.g., Dropped in a specific cell or within a specific boundary. * Option 2: By using a scenario/configuration specific to the local region.   + - E.g., Indoor-outdoor ratio, LOS-NLOS ratio, TXRU mapping, etc.   Note: While modelling the spatial correlation, strive to ensure that the dataset distribution also correctly captures the decorrelation due to temporal variations in the channel. To report methods to generate training and testing dataset. |

Further, in In RAN1#118 meeting [4], Case 2 and Case 3 have been prioritized:

**Agreement**

For temporal domain aspects of AI/ML-based CSI compression using two-sided model in Release 19, among Cases 1, 2, 3, 4, and 5, prioritize further discussion on Case 2 and Case 3.

Based on the aforementioned agreements, we generated datasets for the UMa scenario using Option-1 UE distribution. We assumed Spatial Consistency Procedure A, with UEs distributed in a local region according to Option-1 as per the agreements.

A total of 500 UEs were deployed per sector, with outdoor UEs moving along a track of approximately 10 meters and indoor UEs covering around 4 meters, resulting in approximately 480 time-correlated CSI samples per UE. Consequently, the total dataset comprised approximately 240K samples. Given the random distribution of UEs within a sector, we divided the dataset into 160K samples for training and 80K samples for testing.

The detailed simulation assumptions are described in Appendix.

## AL/ML model Details:

We utilize a Transformer architecture to compress the SVD-based precoder (rank = 1) input at the UE and reconstruct it at the NW side. Both LSTM blocks and multi-head attention (MHA) architectures are used at the encoder and decoder sides to process the information at the UE and NW, respectively. The input data point, Xi , is constructed from 32 antenna ports and spans 32 sub-band frequencies. Using MHA reduces the parameter count while delivering similar performance compared to the LSTM-based architecture.

We evaluate the models for Case 0, Case 2 using the LSTM-based architecture, and for Case 0 and Case 2 using the MHA-based architecture. Our results are compared with Benchmark 2, which corresponds to the AI/ML model evaluated on the spatial-frequency dataset for Case 0.

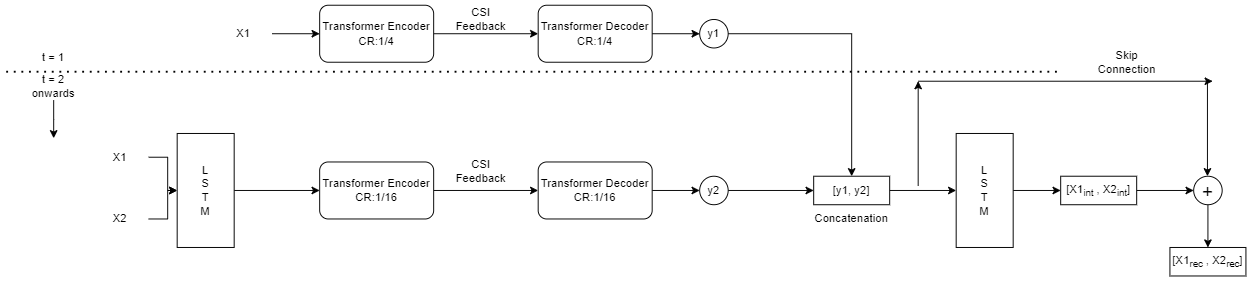


Figure 1: Overview of Model architectures using LSTM

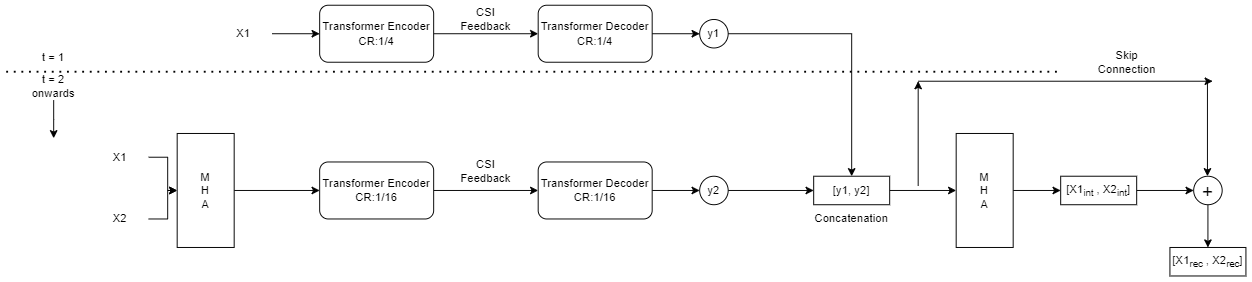


Figure 2: Overview of Model architectures using MHA

## Results

|  |  |  |  |
| --- | --- | --- | --- |
|  | Case 0 | Case 2  (LSTM based arch.) | Case 2  (MHA based arch) |
| Compression  Ratio | 1/16 | 1/16 | 1/16 |
| NMSE (in dB) | -16.55 | -20.26 | -19.56 |
| SGCS | 0.92 | 0.99 | 0.99 |
| Absolute Gain over Case 0 in %  NMSE /(SGCS) | - | 22.4 % / (7.6 %) | 18.18% / (7.6 %) |
| Flops/M | 81.2 | 162.4 | 279.9 |
| Parameters/M | 3.02 | 142.4 | 66.8 |
| Encoder  Parameters/M | 1.51 | 71.2 | 33.4 |
| Decoder  Parameters/M | 1.51 | 71.2 | 33.4 |
| Epochs | 400 | 400 | 400 |

**Observation 1: Utilizing the TSF dataset and incorporating past samples at both the encoder and decoder results in a 7.6% performance SGCS gain over Benchmark 2 for both the LSTM and MHA-based architectures with a payload of 1/16 Compression Ratio. Additionally, the MHA-based architecture achieves comparable performance to the LSTM architecture but with a reduced parameter count.**

It’s important to note that we assume no UCI loss, corresponding to Scenario A. This serves as the benchmark for further evaluation of 10% UCI losses in Scenarios B and C.

# Inter-vendor training collaboration

During the RAN1#116bis meeting, there were many agreements on inter-vendor training collaboration. It was determined that Option 1 can simplify the complexities of inter-vendor collaboration, while Option 2 has been deprioritized. Discussions are ongoing regarding various sub-options for Options 3, 4, and 5.

|  |
| --- |
| * For Option 3, further define the two sub-options:   + 3a: Parameters received at the UE or UE-side goes through offline engineering at the UE-side (e.g., UE-side OTT server), e.g., potential re-training, re-development of a different model, and/or offline testing.   + 3b: Parameters received at the UE are directly used for inference at the UE without offline engineering, potentially with on-device operations. * For Option 5, further define the two sub-options:   + 5a: Model received at the UE or UE-side goes through offline engineering at the UE-side (e.g., UE-side OTT server), e.g., potential re-training, re-development of a different model, and/or offline testing.   + 5b: Model received at the UE are directly used for inference at the UE without offline engineering, potentially with on-device operations. * For Option 4, it is clarified that:   + Dataset received at the UE or UE-side goes through offline engineering at the UE- side (e.g., UE-side OTT server), e.g., model training or offline testing. * For Option 3/4/5, focus further discussion on the following assumptions:   + Option 3a/5a     - The model(5a)/parameter(3a) exchange originates from the NW-side and ends at the UE-side.     - Model(5a)/parameters(3a) exchanged from the NW-side to UE-side is either CSI generation or reconstruction part or both.       * Option 3a-1/5a-1: Model/Parameters exchanged from the NW-side to UE-side is CSI generation part.       * Option 3a-2/5a-2: Model/Parameters exchanged from the NW-side to UE-side is CSI reconstruction part.       * Option 3a-3/5a-3: Model/Parameters exchanged from the NW-side to UE-side are both CSI generation part and CSI reconstruction part.       * Some additional information, if necessary, may be shared from the NW-side to help UE-side offline engineering and provide performance guidance.         + Performance target         + Dataset or information related to collecting dataset     - Study different methods of exchanging, e.g., over the air-interface, offline delivery, etc.   + Option 3b     - The method of exchanging is over the air-interface via model transfer/delivery Case z4.     - The parameter exchange is from NW to UE.     - Parameters exchanged from the NW-side to UE-side is CSI generation part.   + Option 5b     - The method of exchanging is over the air-interface via model transfer/delivery Case z4, assuming that the model structure is aligned based on offline inter-vendor collaboration.     - The model exchange is from NW to UE.     - Model exchanged from the NW-side to UE-side is CSI generation part.   + Option 4:     - The dataset exchange originates from the NW-side and ends at the UE-side.     - Option 4-1: Dataset exchanged from the NW-side to UE-side consists of (target CSI, CSI feedback).     - Option 4-2: Dataset exchanged from the NW-side to UE-side consists of (CSI feedback, reconstructed target CSI).     - Option 4-3: Dataset exchanged from the NW-side to UE-side consists of (target CSI, CSI feedback, reconstructed target CSI).     - Some additional information, if necessary, may be shared from the NW-side to help UE-side offline engineering and provide performance guidance.       * Performance target     - Study different methods of exchanging, e.g., over the air-interface, offline delivery, etc. |

In the RAN1#117 meeting [5], additional agreements were made about inter-vendor training collaboration for AI/ML-based CSI compression. These agreements include the following:

|  |
| --- |
| Conclusion RAN1#117  Standardized signalling, if feasible and specified, can be used for parameter / model exchange in option 3a/5a and 3b to alleviate/resolve the inter-vendor training collaboration complexity.   * Standardized signalling may be reused for exchanging CSI generation part, CSI reconstruction part, or both, etc, when necessary and feasible. * Standardized signalling may be over-the-air, or other approaches.   Standardized signalling, if feasible and specified, can be used for dataset exchange in option 4 to alleviate/resolve the inter-vendor training collaboration complexity.   * Standardized signalling may be reused for dataset exchanging, when necessary and feasible. * Standardized signalling may be over-the-air, or other approaches.   Note: feasibility will be discussed separately.  Agreement RAN1#117   * For option 3a/3b/4/5a and their sub-options, at least the following potential specification impacts have been identified. Further study the necessity, feasibility, their specification impact. * Exchange   + Parameter / model exchange methods, format/contents, and related spec impacts (3a/3b/5a)   + Dataset exchange methods, format/type/contents of data/dataset, and related spec impacts (4)   + Additional information, if necessary, that may be shared from the NW-side to help UE-side offline engineering and provide performance guidance (3a/5a/4)     - Performance target (3a/5a/4)     - Dataset or information related to collecting dataset (3a/5a)     - Any other additional information * Model pairing (3a/3b/4/5a) * UE capability (3a/3b/4/5a) * Model related aspects, such as scalability (e.g., payload sizes, antenna ports, bandwidth), rank and layer handling (3a/3b/4/5a) * Quantization of feedback (3a/3b/4/5a) * Model structure details (3a/3b)   Note: Option 3a/4/5a and option 3b serve two different deployment time scales, UE capabilities, device-side optimizations, and training methods, and therefore may be complementary to each other, with potential specification of both.   * Specification of option 1, if needed from RAN1, can reuse specification of option 3a/3b, with the additional specification of parameters. |

In RAN1#118, it was noted that while Option 5 does not entirely resolve the complexities of inter-vendor collaboration, it offers performance advantages and greater flexibility when compared to Option 3.

Additionally, in the same RAN1#118 meeting [4], the remaining four inter-vendor collaboration options and their corresponding sub-options were consolidated into three distinct approaches, based on on-device operation and offline engineering on the UE side. The potential issues associated with each approach were identified as follows:

|  |
| --- |
| Agreement RAN1#118  Continue the study of following directions - on-device operation and UE side offline engineering   * Direction A: Sharing parameters/reference model/dataset that enables UE-side offline engineering (Inter vendor collaboration option 3a/5a/4)   + Potential down-selection into one or more among sub-options 3a/5a-1, 3a/5a-2, 3a/5a-3, 4-1, 4-2, and 4-3 considering their feasibility and performance, including at least the following issues     - [Issue 1] What additional information should be shared from NW-side to UE-side to enable UE-side encoder training, validation, and testing?     - [Issue 2] Is there concern for NW’s proprietary information disclosure, and if so, how to address it?     - [Issue 3] Is there an overhead concern, and if so, how to address it?     - [Issue 4] Is there performance impact due to mismatch between NW side data distribution and UE side data distribution, and if so, how to address it? * Direction B: Sharing NW side encoder parameter to UE side for UE side inference directly with on-device operation (Inter vendor collaboration option 3b), including at least the following issues   + [Issue 3] Is there an overhead concern, and if so, how to address it?   + [Issue 5] Whether it is feasible to use a common encoder across UEs, and whether it is feasible for NW-side to train multiple encoders for different UEs?   + [Issue 6] Is there performance impact due to mismatch between NW side data distribution and UE side inference data distribution, and if so, how to address it?   + [Issue 7] Is there concern for NW’s and UE’s proprietary information disclosure, and if so, how to address it? * Direction C: Fully standardized reference model(s) and parameters with specified CSI generation part and/or CSI reconstruction part (Inter vendor collaboration option 1), including at least the following issues   + [Issue 8] Whether to consider 3GPP’s statistical channel model or field data for reference model(s) training. In case of the latter, how does RAN1 collect field data and agree on them?   + [Issue 9] Is there performance impact due to mismatch between the distribution of the dataset used for reference model(s) training, UE-side data distribution, and NW-side data distribution, and if so, how to address it?   + [Issue 10] What additional information should be specified to enable UE-side encoder training, validation, and testing, and NW-side decoder training, validation, and testing?   + Note:     - 1-1: Only reference encoder is specified, and NW-side and/or UE-side may train their actual CSI generation part and actual CSI-reconstruction part separately compatible to the reference encoder.     - 1-2: Only reference decoder is specified, and NW-side and/or UE-side may train their actual CSI generation part and actual CSI-reconstruction part separately compatible to the reference decoder.     - 1-3: Both reference encoder and reference decoder are specified, and NW-side and/or UE-side may train their actual CSI generation part and actual CSI-reconstruction part separately that are compatible to the reference decoder/encoder. * Note: UE-side data and NW-side data in “UE-side data distribution” and “NW-side data distribution” are field data. * Note: Some issues identified in one direction may/may not be applicable for other Direction. * Note: potential down selection among the 3 directions is not precluded * Study of data distribution mismatch to consider the use of synthetic data and/or field data. |

For Option 3, when it comes to standardizing the reference model, the work RAN4 is doing on model standardization could be reused if the objectives align, as Option 3 requires a standardized reference model with a known structure. However, when it comes to exchanging parameters, offline methods still leave inter-vendor collaboration issues unresolved, compared to using air-interface signalling. Additionally, CSI reconstruction is expected to play an increasingly important role in overall system performance than CSI generation.

Regarding Issue 2, sub-option 3a poses minimal proprietary concerns on the network side. On the other hand, sub-options 5a-1, and especially 5a-2 and 5a-3, raise concerns because the reconstruction-related information—such as model structure and parameters—could potentially be used to reverse-engineer the system. Additionally, for Options 3a and 5a, the retraining process could introduce delays, which would need to be studied to determine whether they are feasible in practice.

**Observation 2: Sub-option 3a-1 presents minimal proprietary concerns for the network, whereas Option 5a, across all its sub-options, introduces such concerns to varying degrees and relatively more for 5a-2 and 5a-3. For Options 3a and 5a, the potential delays caused by retraining need to be carefully evaluated to see if they are feasible in real-world applications.**

**Proposal 1: Prioritize sub-option 1 for Options 3, 4, and 5—specifically, 3a-1 and 5a-1—over the other sub-options.**

For sub-options 3a-1, 5a-1, and others, preventing performance degradation requires avoiding data drift between the training datasets at the UE and the NW. To address this, the NW would need to share certain information related to the training and testing datasets. However, sharing the entire dataset in real time is not feasible. Instead, the NW could set performance targets for the UE-side training to achieve, particularly in the case of Options 3a, 4, and 5a. In addition, the NW could provide some dataset-related information or a small portion of the dataset to allow the UE to check for any data drift by correlating it with its own dataset.

This approach would require significant standardization efforts, as the UE's performance would need to meet the set targets, and this performance would have to be communicated back to the NW for validation. Based on the computed/reported performance metric, the system may need to fall back to legacy codebooks.

**Proposal 2: For Options 3a and 5a, the NW could share performance targets and/or a portion of the dataset to facilitate UE-side offline engineering and re-training, if necessary. Prioritize a specification impact analysis study for this scenario.**

For Option 4, since the dataset is standardized, there must be alignment between the UE (encoder side) and the NW (decoder side) to ensure proper inference and consistent training on similar structures. This will require some level of standardization effort. Additionally, the UE will need to train its model to meet the performance targets set by the NW, which could introduce delays in retraining.

Additionally, sharing the entire dataset, which is likely to be quite large, could introduce significant overhead. It’s worth studying whether sharing the full dataset and retraining to meet the performance targets is even feasible in real-world scenarios. However, Option 4 has the potential to outperform both Option 1 and Option 3 in terms of performance.

**Observation 3: Option 4 would incur re-training delays and large exchange overhead but has the potential for better performance than Option 1 or 3.**

Regarding Direction B, if the NW-side encoder shares its parameters for direct inference, it would raise minimal proprietary concerns, as no offline engineering or retraining would be necessary. This approach would also reduce signalling overhead and eliminate retraining delays. However, the dataset used to train the NW encoder must accurately reflect the conditions experienced by the UEs, requiring the NW to continuously monitor performance and make adjustments accordingly. Additionally, multiple encoders would need to be trained to accommodate UEs with different capabilities. For UEs with similar capabilities, a common encoder could be used, minimizing the number of encoders the NW needs to train.

In contrast, Direction C would require substantial standardization efforts. While Option 1 offers limited performance compared to Options 3, 4, and 5, it would be more practical to focus on studying the specification impact of both Directions B and C, given the extensive standardization work involved.

**Proposal 3: Prioritize Directions A and B over Direction C, as they provide greater performance potential while requiring less standardization effort.**

# Conclusions

In this contribution, we discussed the two-sided AI/ML model for CSI compression and shared our views on inter-vendor collaboration efforts. The observations and proposals are summarized as follows:

**Observation 1: Utilizing the TSF dataset and incorporating past samples at both the encoder and decoder results in a 7.6% performance SGCS gain over Benchmark 2 for both the LSTM and MHA-based architectures with a payload of 1/16 Compression Ratio. Additionally, the MHA-based architecture achieves comparable performance to the LSTM architecture but with a reduced parameter count.**

**Observation 2: Sub-option 3a-1 presents minimal proprietary concerns for the network, whereas Option 5a, across all its sub-options, introduces such concerns to varying degrees and relatively more for 5a-2 and 5a-3. For Options 3a and 5a, the potential delays caused by retraining need to be carefully evaluated to see if they are feasible in real-world applications.**

**Observation 3: Option 4 would incur re-training delays and large exchange overhead but has the potential for better performance than Option 1 or 3.**

**Proposal 1: Prioritize sub-option 1 for Options 3, 4, and 5—specifically, 3a-1 and 5a-1—over the other sub-options.**

**Proposal 2: For Options 3a and 5a, the NW could share performance targets and/or a portion of the dataset to facilitate UE-side offline engineering and re-training, if necessary. Prioritize a specification impact analysis study for this scenario.**

**Proposal 3: Prioritize Directions A and B over Direction C, as they provide greater performance potential while requiring less standardization effort.**

# References

1. RP-242399, Revised WID on Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface, Qualcomm, RAN#105, Melbourne, Australia, September 9-12, 2024.
2. 3GPP TR 38.843 Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface (Release 18), V2.0.0 (2023-12)
3. R1- 2403505 Final summary of Additional study on AI/ML for NR air interface: CSI compression, Moderator (Qualcomm), 3GPP TSG RAN WG1 #116-bis, Changsha, China, April 15th – April 19th, 2024
4. Chair’s Notes RAN1#118, Maastricht, NL, August 19th – 23th, 2024
5. Chairman's Notes RAN1#117, Fukuoka City, Fukuoka, Japan, May 20th – 24th, 2024.

# Appendix

Evaluation Assumptions:

|  |  |
| --- | --- |
| Carrier Frequency | 2GHz (FR1) |
| SCS | 15Khz |
| Simulation bandwidth | 12 MHz |
| Scenario | Uma |
| Num of RB | 32 |
| Num\_gNB | 1 |
| Num\_sectors | 1 |
| gNB\_height | 25 m |
| gNB Antenna config | 32 ports (8,2,2,1,1) |
| UE Antenna config | (1,1,2,1,1) |
| Num\_UEs | 500 |
| Tx power | 40 dBm |
| UE receiver noise figure | 10 dB |
| Cell Radius | 100 m |
| UE Distribution | 1) 80% indoor, 20% outdoor |
| UE\_speed | Outdoor: 10Km/hr, Indoor: 3 Km/hr |
| UE Track Length | 10 m, 4m |
| CSI-RS periodicity | 5 ms |
| CSI reporting periodicity | 5 ms |
| Spatial consistency | Procedure A (3GPP 38.901) |
| Number, distance of past time instance | 1, 5 ms |