**3GPP TSG-RAN WG1 Meeting #118 R1-** **24xxxxx**

**Maastricht, Netherland, August 19th – August 23th, 2024**

**Agenda item:** 9.1.3.2

**Source:** Moderator (Qualcomm)

**Title:** Draft summary of Additional study on AI/ML for NR air interface: CSI compression

**Document for:** Discussion and Decision

# Introduction

In RAN#102 plenary meeting, a new WID on Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface was approved ‎[2]. The WID includes study objectives related to AI/ML for CSI compression using a two-sided model.

|  |
| --- |
| ……  Study objectives with corresponding checkpoints in RAN#105 (Sept ’24):   * 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 other aspects requiring further study/conclusion as captured in the conclusions section of the TR 38.843.   * + ……   …… |

This document summarizes the issues regarding agenda item 9.1.3.2 (Additional study on AI/ML for NR air interface: CSI compression) in RAN#118.

## Contact Information

Please provide / update your contact information.

|  |  |  |
| --- | --- | --- |
| **Company** | **Name** | **E-mail** |
| Moderator (Qualcomm) | Taesang Yoo | taesangy@qti.qualcomm.com |
| Lenovo | Vahid Pourahmadi | vpourahmadi@lenovo.com  wangjf20@lenovo.com |
| New H3C | Lei Zhou | Zhou.leih@h3c.com |
| Samsung | Ameha Tsegaye Abebe | amehat.abebe@samsung.com |
| Indian Institute of Technology Madras | Anil Kumar Yerrapragada | anilkumar@5gtbiitm.in |
| Fujitsu | WANG Guotong (David) | wangguotong@fujitsu.com |
| vivo | Peng SUN | sunpeng@vivo.com |
| SK telecom | Yunesung Kim | yunesung.kim@sktelecom.com |
| Mavenir | Ali Fatih Demir | ali.demir@mavenir.com |
| Panasonic | Tetsuya Yamamoto  Hidetoshi Suzuki | [yamamoto.tetsuya001@jp.panasonic.com](mailto:yamamoto.tetsuya001@jp.panasonic.com)  suzuki.hidetoshi@ jp.panasonic.com |
| OPPO | Wendong Liu | liuwendong1@oppo.com |
| NTT DOCOMO | Xin Wang  Haruhi Echigo | [wangx@docomolabs-beijing.com.cn](mailto:wangx@docomolabs-beijing.com.cn)  haruhi.echigo.fw@nttdocomo.com |
| ETRI | Anseok Lee | alee@etri.re.kr |
| Ericsson | Jingya Li  Siva Muruganathan  Jianwei Zhang | [Jingya.li@ericsson.com](mailto:Jingya.li@ericsson.com)  [Siva.muruganathan@ericsson.com](mailto:Siva.muruganathan@ericsson.com)  Jianwei.zhang@ericsson.com |
| ZTE | Lun Li  Xingguang Wei | [li.lun1@zte.com.cn](mailto:li.lun1@zte.com.cn)  wei.xingguang@zte.com.cn |
| IIT Kanpur | Shyam Vijay Gadhai  Abhishek Kumar Singh | [svgadhai@iitk.ac.in](mailto:svgadhai@iitk.ac.in)  [abhishekks@iitk.ac.in](mailto:abhishekks@iitk.ac.in) |
| Spreadtrum | Mimi Chen | Mimi.chen@unisoc.com |
| LG | Jaehoon Chung | jhoon.chung@lge.com |
| CATT | Qianrui Li | liqianrui@catt.cn |
| Intel | Victor Sergeev | Victor.sergeev@intel.com |
| Xiaomi | Liuzhengxuan  Liumin | liuzhengxuan@xiaomi.com  liumin10@xiaomi.com |
| AT&T | Isfar Tariq  Salam Akoum | [Isfar.tariq@att.com](mailto:Isfar.tariq@att.com)  [Salam.akoum@att.com](mailto:Salam.akoum@att.com) |
| CMCC | Yuhua Cao  Yi Zheng  Dan Song | [caoyuhua@chinamobile.com](mailto:caoyuhua@chinamobile.com)  [zhengyi@chinamobile.com](mailto:zhengyi@chinamobile.com)  songdan@chinamobile.com |
| NVIDIA | Xingqin Lin | [xingqinl@nvidia.com](mailto:xingqinl@nvidia.com) |
| MediaTek | Pedram Kheirkhah Sangdeh | Pedram.kheirkhah@mediatek.com |
| Futurewei | Baoling S Sheen | [bsheen@futurewei.com](mailto:bsheen@futurewei.com) |
| TCL | Yunsheng Kuang | yunsheng.kuang@tcl.com |
| Huawei, HiSilicon | Yuan Li | liyuan3@huawei.com |
| CEWiT(TSDSI) | Shiv Shankar  Dhivagar Baskaran | [shivshankar@cewit.org.in](mailto:shivshankar@cewit.org.in)  [dhivagar.b@cewit.org.in](mailto:dhivagar.b@cewit.org.in) |
| NEC | Zhen He  Pravjyot Deogun | he\_zhen@nec.cn  pravjyot.deogun@EMEA.NEC.COM |
| Qualcomm | Chenxi Hao | [chenxih@qti.qualcomm.com](mailto:chenxih@qti.qualcomm.com) |
| Sony | Hiroki Matsuda  Sam Atungsiri | [hiroki.matsuda@sony.com](mailto:hiroki.matsuda@sony.com)  sam.atungsiri@sony.com |
| Apple | Huaning Niu | Huaning\_niu@apple.com |
| Tejas Networks | Pavan Kalyan | [pavankalyand@tejasnetworks.com](mailto:pavankalyand@tejasnetworks.com) |

# FL’s questions and comments to Companies regarding results template

Please provide clarifications on the following questions regarding what you captured in the results template.

**If you make any corrections in the Results Template, be sure to**

* **Notify your correction by adding comments in the table below**
* **Use red colour for any updates in the Results Template**

## CSI\_Table X1 (Cases 1,2,5)

* To Nokia: FLOPs are missing for case 2.
* To Samsung: Please verify that the model complexity is 0.2 FLOPs for case 2. Also, columns (AN, AO,AP) are duplicates and looks like they represent same results in column (M,N,O). The results in new columns are shifted by one row, which looks like a mistake.
* To CATT: Please specify the FLOPs for CSI generation for case 2 (nothing on Excel and Word doc indicates scattered from 10M-800M).
* To CATT: Mismatch between Excel and Word tdoc for the SGCS values.
* To OPPO: FLOPs and number of parameters have been specified same. Please confirm.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Samsung | Thank you FL. It is right, the 3 columns (AN, AO, AP) are duplications of (M,N,O). Now, deleted from the Results Template. |
| OPPO | Thanks to FL. We have some mistakes on our FLOPs results. The FLOPs should be 277.9M for both encoder and decoder. The number of parameters are still 10.7M. |

## CSI\_Table X2 (Cases 3,4)

* To Samsung, Fujitsu, ZTE, Qualcomm, CATT, CMCC, Nokia, InterDigital: For the “FLOPs/M and FLOPs/M/5msec” row, you only listed one number. I assumed that the number you provided is FLOPs/M and not FLOPs/M/5msec. Please correct the spreadsheet and notify your correction by adding comments in the table below, in case my assumption is wrong.
* To Samsung: baseline results in the case3 spreadsheet is for Rel-16 etypeII. Please provide the agreed baseline Rel-18 etypeII if you have.
* To Oppo: FLOPs/M and FLOPs/M/5mse, and number of parameters are missing for CSI generation and reconstruction.
* To Nokia: baseline and SGCS % gain is missing for case 4.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Samsung | Thank you FL. It is FLOPs/M. |
| OPPO | Thanks to FL. We missed this item. For FLOPs/M, about 1000M for CSI generation and reconstruction. For FLOPs/M/5mse, about 278M for CSI generation and reconstruction. The number of parameters is 10.7M for CSI generation and reconstruction. |
| ZTE | According to the submitted results in Table X2, we have already splitted the “FLOPs/M” and “FLOPs/M/5msec” into two rows, which are shown in the 1st row and 2nd row, respectively. Please FL further check our results. |

## CSI\_Table X3 (Localized Models):

* To Intel: In the full buffer results section for Localized Model study for cell edge throughputs in the “A” bin, we had a question. With a low complexity local model, the reported gain over global model is 3.4%, while the reported loss over baseline eType2 is 7% (-7% gain). The high complexity local model does only 2.2% better than global model and only 2% worse than R16 eType2. So, the low complexity local model is doing much worse against eType2 than the global model is, but is simultaneously better than high complexity local model. Just wanted to verify if the correct numbers have been copied into the spreadsheet, or if our understanding of the result is faulty.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Intel | Thanks to the FL for checking with us.  Our understanding for the value corresponding to gain comparing to global model was to look into gains provided only by site-specific training with fixed complexity. So, the value provided by us is the gain for low complexity site-specific model over low complexity global model (the difference is local/global dataset).  Indeed, if we compare with high complexity global model, the gains for low complexity local model are:  **Full buffer**   * **Mean UPT**: -1.2% * **Cell-edge UPT**: -3.0%   **FTP**   * **Mean UPT**: -0.1% / -0.5% / -1.4% * **Cell-edge UPT**: 0.2% / -2.4% / -4.1% |
|  |  |

## Others

Please provide any other comments for this section.

|  |  |
| --- | --- |
| *Company* | *Comments* |
|  |  |
|  |  |

# Temporal domain aspects of CSI compression

The Release 19 work item description ‎[2] has listed improving the trade-off between performance and complexity/overhead as one of the study objectives and has provided several example approaches. This section discusses the aspects of “extending the spatial/frequency compression to spatial/temporal/frequency compression” and “CSI compression plus prediction”. In this document, the term “temporal domain aspects of CSI compression” is used as a general term to refer to both these aspects.

## Summary of company proposals

### Case 1/2/5

Huawei, HiSilicon

***Proposal 1: For the evaluation of non-ideal UCI feedback in Case 2, Case 4, and Case 5, it can be modelled with a missing rate (e.g., 10%) for each individual CSI report occasion.***

***Proposal 2: For the additional potential spec impact of temporal domain CSI compression Case 2 on top of Rel-18 SF domain CSI compression, consider methods to handle the misalignment of the accumulated CSI between NW part model and UE part model due to UCI missing.***

Spreadtrum Communications, BUPT

***Proposal 2: Consider reporting historical CSI information via NW-triggered signaling when UCI missing or UCI dropping.***

Tejas Networks

***Proposal 1: For the EVM of temporal domain CSI compression, consider the following assumptions for the CSI generation part and CSI reconstruction part, respectively:***

* ***CSI generation part at t=T:***
* ***Model input: Pre-processed channel matrix is given as a input to the CsiNet encoder.***
* ***First layer of encoder is Convolutional layer with real and imaginary parts of channel as its input. This layer uses 33 kernel size to generate two feature maps. Following the convolutional layer, we reshape the feature maps into a vector and use a fully connected layer to generate the codeword s, which is a real-valued vector of size M.***
* ***Model output: Encoder transforms channel matrix into a vector of size M.***
* ***CSI reconstruction part at t=T+∂ (where ∂ is an uplink latency)***
* ***Model input: Codeword of size Mis given as a input to the CsiNet decoder.***
* ***Model output: First layer of decoder is dense fully connected layer which gives the initial estimate of the channel followed by several RefineNet units gives the CSI.***

***FFS: Study the effect of quantization***

**Proposal 2: Consider *a missing rate (e.g., 10%) for each individual CSI report occasion to model UCI loss in case 1 and case 5.***

Intel

***Proposal 1***:

* ***For the observations with AI/ML CSI compression performance gains over PMI codebook, capture range for the AI/ML model complexity together with the corresponding range for performance gains.***
  + ***For all the observations with AI/ML model complexity, the following note should be added: “AI/ML model complexity depends on various platform-dependent choices for model implementation. The values reported here should be considered as representative values and not as a precise complexity estimate for implementation of AI/ML-based CSI compression.”.***

ZTE

***Proposal 4: For the evaluation of non-ideal UCI feedback in Case 2, Case 4, and Case 5, it can be modelled with a missing rate (e.g., 10%) for each individual CSI report occasion.***

***Proposal 5: For* *temporal domain CSI compression Case 2, further study the performance impact resulting from the aperiodic CSI feedback.***

OPPO

*Proposal 1: For the evaluation of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Rel-19, suggest to down-select from Case 0 - Case 5:*

* ***Study Case 2 without CSI prediction in high priority***
* ***Study Case 3 with CSI prediction in high priority***
* ***Study Case 1/4/5 in low priority***

***Note: Companies report how the past CSI information is used in different cases.***

*Proposal 2: Regarding different training types for AI/ML-based CSI compression using two-sided model with temporal domain CSI correlation, suggest:*

* ***Type 1 and Type 3 should be treated in priority***
  + ***Evaluations on Type 1 should be firstly considered***
  + ***Type 3 related issues, e.g., temporal information indicating, alignment of past CSI information utilization, can be discussed in parallel***
* ***Type 2 is deprioritized***

*Proposal 3: Regarding the training and deploy methodology of SFT-domain CSI compression, two kinds of assumptions can be considered:*

* ***Assumption 1: with time window (baseline)***
* ***Assumption 2: without time window (optional)***
  + ***How to perform model training under Assumption 2 should be studied***

***Note: Companies to report which assumption is selected.***

*Proposal 4: Regarding the modeling assumption of non-ideal UCI feedback, use Assumption 1 for evaluations. (Observation 6: Regarding the modeling of non-ideal UCI feedback, two kinds of modeling assumptions can be considered:*

* ***Assumption 1: UCI loss happens in p% probability for each slot of CSI feedback***
* ***Assumption 2: UCI reporting error in p% probability for each slot of CSI feedback****)*

*Proposal 5: Suggest no further evaluation and discussion on Case 1*

*Proposal 6: Regarding the model of SFT-domain CSI compression, a proper time window size is required to achieve the trade-off between performance and complexity*

Xiaomi

***Proposal 14: Recommend the two-sided AI/ML model based CSI compression to study as a normative work, and at least Case 2 and Case 3 should be supported.***

Fujitsu

***Proposal 1:***

* ***For the study of the performance impacts resulting from UCI loss, the following two options could be considered as a starting point for Case 2 if UCI loss happens:***
  + ***Option A: Past CSI information is reset at NW side only.***
  + ***Option B: Past CSI information is reset at both UE and gNB sides.***

***Proposal 2:***

* ***For the evaluation of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Release 19, for Case 2/5, RAN1 to discuss how to address the issue of unavailable past CSI information for some layers resulting from rank adaption, e.g., how to reset the past CSI information, and study the performance impact.***

***Proposal 3:***

* ***Regarding the use cases for CSI feedback enhancement studied in Rel-19, if AI/ML based CSI compression with two-sided model will be specified in Rel-19, it is recommended to consider Case-2 and Case-3 for normative work based on the evaluation results***
  + ***Case-2 could be considered for indoor scenario***
  + ***Case-3 could be considered for both indoor and outdoor scenario***
* ***If the time is not sufficient to specify AI/ML based CSI compression within Rel-19 timeframe, further evaluation and study could be considered in Rel-19.***

CATT

**Proposal 1: If AI/ML-based CSI compression is specified in the WI phase in Rel-19, SF-domain AI/ML-based CSI compression is selected as the basic feature.**

LG Electronics

**Proposal #1: Regarding temporal/spatial/frequency (TSF)-domain CSI compression, study methods/mechanisms to manage the similarity/synchronization of accumulated past CSI at UE-side and/or NW-side.**

**Proposal #2: Regarding TSF-domain CSI compression, discuss the format of past CSI information and how to report it at least for performance monitoring perspective.**

**Proposal #4: Regarding non-ideal UCI feedback on TSF-domain CSI compression,**

* **Consider two-step performance monitoring to check that the performance degradation of the AI/ML model is originated from whether the accumulated past CSI has a problem or the AI/ML model is not suitable for the deployed environment**
* **Also consider to report past CSI information via NW-triggered signaling when UCI missing or UCI dropping.**

Lenovo

***Proposal 23: Prioritize Case 2 and Case 3 for temporal domain aspects of AI/ML-based CSI compression using two-sided model.***

***Proposal 24: Support procedures/signalling enabling CSI-compression models having both Scaler and vector Quantizers for generation of the CSI-feedback bits.***

InterDigital, Inc.

**Proposal 3: TSF compression performance should be evaluated under multiple observation window lengths.**

**Proposal 4: For AI-TSF compression Case2 with missing UCI, study different missing UCI mitigation solutions, e.g., buffer reset, to handle error propagation.**

**Proposal 5: AI/ML CSI compression should continue as a study item for the remainder of Rel-19.**

NEC

***Proposal 1: RAN1 to prioritize to study Case 2 and Case 3.***

***Proposal 8: At least for Case 2, CSI buffer reset should be supported to address*** ***misalignment of historical CSI used at UE side and NW side. And the definition, determination or indication of the reset value need to be further studied.***

***Proposal 9: At least for Case 2, further study*** ***effective availability of historical CSI information over time.***

Nokia

Proposal 1: Develop standardized mechanisms to address synchronization of state for recurrent models used for CSI compression.

Proposal 2: Consider exploiting the time-domain (Case 2) by using recurrent quantization / inverse quantization together with history-independent SF based dimensionality reduction, instead of using SFT-based encoder / decoder pairs.

Samsung

Proposal#1: Among the identified six cases for AI/ML-based CSI compression using two-sided model, deprioritize Case 1 as its additional spec. impact compared to Case 0 is not clear.

Proposal#2: Among the identified six cases for AI/ML-based CSI compression using two-sided model, for Case 2, consider at least the following two options for the past CSI information

•Case 2-1: Past CSI information generated by the UE-part and/or network-part of two-sided model

•Case 2-2: Information on SD/FD basis vectors as past CSI information with angle-delay domain compression.

Proposal#3: Among the identified six cases for AI/ML-based CSI compression using two-sided model, for Case 2:

•when past CSI information corresponds to SD/FD basis and AI/ML CSI compression in the angle-delay domain, consider SD/FD basis reporting per N CSI reporting occasions, i.e., N times longer periodicity.

•FFS on the values of N.

Proposal#5: For cases that utilize past CSI reports at the network (Case 2/4/ 5), RAN1 to study the error propagation that may result from

- Imperfect past CSI generation ( representation)

- Part 1 and/or Part II CSI dropping (depending on priority)

- UCI transmission loss

Proposal#6: For cases that utilizes past CSI reports at the network (Case 2/4/ 5), for the evaluation of the error propagation

- The impact from imperfect past CSI generation ( representation) can be inherently captured

- Consider different dropping probabilities for the priority levels of part 1 and part II CSI with a decreasing probability as priority increases.

- Consider a fixed UCI transmission loss probability of a CSI report

Note: Companies to report the partitioning of part I and part II CSI

Proposal#7: Study the impact of input pre-processing (dimensionality reduction) on performance and model complexity.

Proposal#8: In Angle-delay (W2)-domain CSI compression, study the impact of the number of SD/FD basis vectors for performance-complexity tradeoff.

ETRI

**Proposal 1: For the study of temporal domain aspects of AI/ML-based CSI compression using the two-sided model in Release-19, select case(s) to prioritize for evaluation and discussions.**

**Proposal 2: For the study of temporal domain aspects of AI/ML-based CSI compression using the two-sided model in Release-19, prioritize evaluations and discussions of Case 2 and 4.**

**Proposal 3: For AI/ML-based CSI compression using two-sided model, when UE and/or NW uses past CSI information, reuse the current specification on CSI-RS transmissions as much as possible.**

**Proposal 4: For AI/ML-based CSI compression using two-sided model, when NW uses past CSI information, study method to detect and mitigate inconsistency of the availability of past CSI information between the UE and the NW.**

Mediatek Inc

1. Evaluate the feedback error tolerance of eType II and compare it with that of AI/ML model.

Apple Inc

**Proposal 2: For case 2 and case 4 of time-frequency-spatial domain CSI compression, the following potential specification impact are proposed:**

* **Enable semi-persistent CSI reporting**
* **Enable DCI based reset state**
* **Considering UCI retransmission in case of large amount of UCI drop or loss, to avoid the state at UE and gNB out of sync.**

**For RI, consider longer term RI update across different CSI reports**

**Proposal 6: For time-frequency-spatial domain CSI compression, flexible CSI report configuration to support different cases should be studied.**

**Proposal 7: For CSI compression using two-sided model, for UE side performance, further study RLF/BFD like mechanism for UE initiated report.**

AT&T

**Proposal 1: Deprioritize case 1/3/4/5 for AI/ML-based CSI compression using two-sided model in Release 19.**

Qualcomm

***Proposal 10: For facilitating inter-vendor collaboration in case 2, support refinement model as a candidate NN design, where all NN components except for the differential quantizer(s) are trained following same methodology in case 0. The quantizers (base/differential) can be exchanged from NW-side to the UE-side.***

### Case 3/4

Huawei, HiSilicon

***Proposal 1: For the evaluation of non-ideal UCI feedback in Case 2, Case 4, and Case 5, it can be modelled with a missing rate (e.g., 10%) for each individual CSI report occasion.***

***Proposal 3: For the additional potential spec impact of temporal domain CSI compression Case 3 on top of Rel-18 SF domain CSI compression,***

* ***for separate prediction and compression, potential spec impact may be needed for data collection of the CSI compression model and monitoring of the CSI prediction model.***
* ***for joint prediction and compression, potential spec impact may be needed for data collection, inference, and monitoring.***

Tejas Networks

**Proposal 3: *In the results template for capturing the evaluation of temporal domain aspects Case 3/4 of AI/ML based CSI compression, it is clarified that the upper bound is calculated based on ideal CSI prediction and without CSI compression.***

**Proposal 4: *For temporal domain aspects Case 3 and 4, study the impact on LCM aspects (e.g., data collection, training, monitoring, and model control) of separate prediction and compression vs. joint prediction and compression.***

ZTE

***Proposal 1: For temporal domain CSI compression Case 3, legacy CSI prediction plus AI CSI compression should be prioritized to study and evaluate the performance.***

***Proposal 2: For legacy CSI prediction plus AI CSI compression sub-use case in Case 3, further study and evaluate at least the following potential case:***

***• Model input: predicted precoding matrices of multiple instances per layer***

***• Model output: recovered predicted precoding matrix of each one instance per layer***

***Proposal 3: For results template Table 1, further adopt the two sets of parameters, i.e., α and β, to capture the evaluation results of high feedback overhead for temporal domain CSI compression Case 3/4 when the length of prediction window is different,***

***• When the length of prediction window is 1, the legacy parameters in Rel-18 evaluation table are reused;***

***• When the length of prediction window is larger than 1, the new parameters proposed in RAN1#117 meeting can be introduced.***

***Proposal 6: For the Rel-19 temporal domain CSI compression cases, at least prioritize the study on temporal domain CSI compression Case 3.***

OPPO

***Proposal 1: For the evaluation of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Rel-19, suggest to down-select from Case 0 - Case 5:***

***• Study Case 2 without CSI prediction in high priority***

***• Study Case 3 with CSI prediction in high priority***

***• Study Case 1/4/5 in low priority***

***Note: Companies report how the past CSI information is used in different cases.***

***Proposal 2: Regarding different training types for AI/ML-based CSI compression using two-sided model with temporal domain CSI correlation, suggest:***

***• Type 1 and Type 3 should be treated in priority***

***- Evaluations on Type 1 should be firstly considered***

***- Type 3 related issues, e.g., temporal information indicating, alignment of past CSI information utilization, can be discussed in parallel***

***• Type 2 is deprioritized***

Xiaomi

***Proposal 12: If multiple predicted CSI of the multiple future instances are reported in one CSI reporting, how to pack the multiple CSI in the CSI reporting needs to study.***

***Proposal 14: Recommend the two-sided AI/ML model based CSI compression to study as a normative work, and at least Case 2 and Case 3 should be supported.***

Fujitsu

***Proposal 3:***

* ***Regarding the use cases for CSI feedback enhancement studied in Rel-19, if AI/ML based CSI compression with two-sided model will be specified in Rel-19, it is recommended to consider Case-2 and Case-3 for normative work based on the evaluation results***
  + ***Case-2 could be considered for indoor scenario***
  + ***Case-3 could be considered for both indoor and outdoor scenario***
* ***If the time is not sufficient to specify AI/ML based CSI compression within Rel-19 timeframe, further evaluation and study could be considered in Rel-19.***

CATT

***Proposal 13: In CSI compression using two-sided model use case, for L1 signalling based reporting of ground-truth CSI/target CSI for NW-side data collection, study whether multiple CSI in the same report is supported for Case 2/3/4.***

***Proposal 20: In CSI compression using two-sided model use case, for temporal domain aspects Case 3 and Case 4 with separate prediction and compression adopted, support monitoring the performance of the model for prediction and the performance of the model for compression separately.***

LG

***Proposal #3: Regarding TSF-domain CSI compression Case 3 and 4, consider performance monitoring method on joint CSI compression and prediction by adapting the operation on the AI/ML model between CSI compression and prediction.***

Lenovo

***Proposal 23: Prioritize Case 2 and Case 3 for temporal domain aspects of AI/ML-based CSI compression using two-sided model.***

NEC

***Proposal 1: RAN1 to prioritize to study Case 2 and Case 3.***

Nokia

***Proposal 3: RAN1 should further study the joint CSI prediction and compression (case 3 and 4) across different scenarios, and compared results to those obtained using separate ML models for CSI prediction and CSI compression.***

Samsung

***Proposal#4: Among the identified six categories for AI/ML-based CSI compression using two-sided model, for Case 3 and Case 4, consider N\_4≥1 prediction instances (Doppler time intervals).***

***Option1: AI/ML-based CSI compression in spatial-frequency-time domain***

***Option 2: The AI/ML-based CSI compression in angle-delay-time domain***

***Option 3: The AI/ML-based CSI compression in angle-delay-Doppler domain***

***Proposal#5: For cases that utilize past CSI reports at the network (Case 2/4/ 5), RAN1 to study the error propagation that may result from***

***- Imperfect past CSI generation ( representation)***

***- Part 1 and/or Part II CSI dropping (depending on priority)***

***- UCI transmission loss***

***Proposal#6: For cases that utilizes past CSI reports at the network (Case 2/4/ 5), for the evaluation of the error propagation***

***- The impact from imperfect past CSI generation ( representation) can be inherently captured***

***- Consider different dropping probabilities for the priority levels of part 1 and part II CSI with a decreasing probability as priority increases.***

***- Consider a fixed UCI transmission loss probability of a CSI report***

***Note: Companies to report the partitioning of part I and part II CSI***

***Proposal#9: For the cases of CSI compression with temporal aspects, consider the following for the network’s ground-truth CSI collection***

* ***For cases that require multiple time-domain samples for inference, cases 2/3/4/5, high resolution codebook quantization including temporal aspects, e.g., Rel-18 eType II-like method with new parameters.***
* ***For cases with CSI prediction, e.g., cases 3/4, high resolution codebook quantization for explicit channel matrices, e.g., codebook to report the left and right eigenvectors of a channel matrix H=UΛV^H***

ETRI

***Proposal 1: For the study of temporal domain aspects of AI/ML-based CSI compression using the two-sided model in Release-19, select case(s) to prioritize for evaluation and discussions.***

***Proposal 2: For the study of temporal domain aspects of AI/ML-based CSI compression using the two-sided model in Release-19, prioritize evaluations and discussions of Case 2 and 4.***

***Proposal 5: For AI/ML-based CSI compression using two-sided model, when the target CSI is Future slot(s), study following aspects, for performance monitoring operations:***

* ***Method to align whether prediction and compression occur in separate steps or simultaneously between UE and NW***
* ***Either UE-side or NW-side performance monitoring***

MediaTek Inc.

***Proposal 1. Deprioritize case 3 and case 4 which may use one AI/ML model for doing both CSI compression and CSI prediction in this agenda item.***

Apple Inc.

***Proposal 2: For case 2 and case 4 of time-frequency-spatial domain CSI compression, the following potential specification impact are proposed:***

***• Enable semi-persistent CSI reporting***

***• Enable DCI based reset state***

***• Considering UCI retransmission in case of large amount of UCI drop or loss, to avoid the state at UE and gNB out of sync.***

***• For RI, consider longer term RI update across different CSI reports.***

***Proposal 4: For performance monitoring of case 4 of time-frequency-spatial domain CSI compression, CSI measurement in prediction window is the target CSI for NW side or UE side performance monitoring. The intermediate KPI or eventual KPI includes both compression and prediction performance.***

***Proposal 5: For case 3 of time-frequency-spatial domain CSI compression, CSI measurement in prediction window is the target CSI for NW side or UE side performance monitoring. The intermediate KPI or eventual KPI includes both compression and prediction performance.***

***Proposal 6: For time-frequency-spatial domain CSI compression, flexible CSI report configuration to support different cases should be studied.***

AT&T

***Proposal 1: Deprioritize case 1/3/4/5 for AI/ML-based CSI compression using two-sided model in Release 19.***

Qualcomm

***Proposal 11: For facilitating inter-vendor collaboration in case 3, consider separate prediction and compression architecture as a baseline design, where training and designing the prediction module is up to UE implementation. NW-side training is performed assuming ideal prediction and inter-vendor collaboration is focused on training the compression model***

***Proposal 12: For facilitating inter-vendor collaboration in case 4, consider separate prediction and compression with refinement architecture as a baseline design, where training and designing the prediction module is up to UE implementation. NW-side training is performed assuming ideal prediction and inter-vendor collaboration is focused on training the compression model. The quantizers (base/differential) can be exchanged from NW-side to the UE-side.***

CEWiT

***Proposal-1: In case of Case-3 and Case-4 based CSI compression, study the effects of having a separate prediction module versus compression plus prediction module at the UE side.***

***Proposal-2: Study methods to model the absence of past CSI in the case of rank adaptation in Case-3 and Case-4 based CSI compression.***

## Discussion

### Prioritization of Temporal Cases

The following table summarizes the number of sources for each of the temporal compression cases, based on the intermediate-KPI data points.

|  |  |  |
| --- | --- | --- |
| **Cases** | **# of sources** | **Sources (Intermediate KPI)** |
| **Case1** | 3 | [Futurewei, OPPO, CMCC] |
| **Case2** | 17 | [Fujitsu, ZTE, Apple, QC, Samsung, vivo, OPPO, Xiaomi, Spreadtrum, Huawei, ETRI, Nokia, Futurewei, CATT, CMCC, IIT-Kanpur, InterDigital] |
| **Case3** | 12 | [oppo, vivo, QC, Fujitsu, ZTE, DOCOMO, CATT, Ericsson, Samsung, CMCC, Xiaomi, InterDigital] |
| **Case4** | 4 | [QC, ETRI, Apple, Nokia] |
| **Case5** | 3 | [Fujitsu, OPPO, IIT-Kanpur] |

Given the limited evaluation results on cases 1/4/5, it is reasonable to focus the discussion on case 2/3 and deprioritize the other cases. Additionally, several companies are proposing the prioritization of case 2 and/or Case 3.

Proposal 1a:

For the evaluation of 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 and study on Case 2 and Case 3.

|  |  |
| --- | --- |
| *Support / Can accept* |  |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Support |
| Xiaomi | Support |
| Huawei, HiSilicon | Support |
| Lenovo | Support |
| NEC | Support |
| Panasonic | Support |
| Futurewei | Support |
| Samsung | Ok |
| ETRI | We can live in the proposal for the progress. |
| SK telecom | Support |
| LG | Fine |
| OPPO | support |
| ZTE | Support |
| Ericsson | Support |
| Spreadtrum | support |

### Separate vs. joint prediction and compression for Case 3

For temporal domain aspects Case 3, the UE may perform prediction as a separate step or jointly with compression. Companies noted that whether the prediction and compression is performed jointly or separately may affect LCM aspects such as data collection, training, monitoring, and model control. So, it is suggested to study this further.

From the evaluation results on Case 3, majority of companies considered separate prediction and compression (SPC) over joint prediction and compression (JPC). Additionally, some companies are proposing the prioritization of SPC.

|  |  |  |
| --- | --- | --- |
| **Case-scheme** | **Companies evaluation** | **Support** |
| Case3-SPC | [oppo, vivo, QC, Fujitsu, ZTE, DOCOMO, CATT, Ericsson, Samsung, CMCC, Xiaomi, InterDigital] | [QC, MediaTek] |
| Case3-JPC | [CMCC] |  |

Proposal 2a:

For temporal domain aspects Case 3, take separate prediction and compression (SPC) as a baseline, and study its LCM aspects and specification impacts. FFS: Joint prediction and compression (JPC), its LCM aspects and specification impacts.

For SPC, study the following options for training data collection

* Option 1: The target CSI for training is derived based on the predicted CSI of the future slot(s).
* Option 2: The target CSI for training is derived based on the measured CSI of the future slot(s).
* Note: During inference, the input to the CSI generation part is derived based on the predicted CSI.

For SPC, study the following options for the monitoring target

* Option 1: The monitoring target is derived based on the predicted CSI of the future slot(s).
  + Note: This corresponds to monitoring of CSI compression only. CSI prediction may be monitored separately.
* Option 2: The monitoring target is derived based on the measured CSI of the future slot(s)
  + Note: This corresponds to end-to-end monitoring of CSI prediction and compression.

For SPC, study how the functionality/model control (activation, deactivation, switching, and fallback) for CSI prediction and CSI compression interacts.

|  |  |
| --- | --- |
| *Support / Can accept* |  |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Support but we think this may not be an urgent issue to discuss for now. |
| Xiaomi | Support to study for this direction. |
| Huawei, HiSilicon | In general, agree with the main text that SPC is considered as a starting point.  For the training data collection, does “The target CSI for training” mean for training the CSI compression model? If so, agree to study the two options. Option 1 seems more reasonable to us as the label for training is more in-line with the input/target for inference.  For the monitoring target, we prefer Option 1, since Option 2 (end-to-end monitoring) is more close to JPC. |
| Panasonic | We are fine with the proposal, but we share the vivo’s view. |
| Futurewei | We agree the proposal in general. If CSI compression is agreed to progress to normative work in Rel-19, we suggest supporting SPC only. |
| Samsung | Same view as vivo. With Option 2 for training and monitoring, there may not be much difference between SPC and JPC. |
| ETRI | For the training data collection, we think Option 2 is a more preferred direction because the eventual target at the output of CSI reconstruction part is the future (measured) CSI.  For the monitoring target, in our view, we need a monitoring target for measuring compression performance and another monitoring target for measuring prediction performance since this is the separate prediction and compression case. We think this point could be captured as another option (Option3?). |
| LG | Fine with the direction |
| OPPO | Same view to vivo. |
| ZTE | Agree with vivo and Samsung, it may not be urgent to discuss now. |
| Ericsson | Support the direction. |

### Non-ideal UCI feedback for Case 2

Proposal 3a:

For the evaluation of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Release 19, for Case 2, Case 4 and Case 5, study the performance impact resulting from non-ideal UCI feedback.

* Scenario A: no UCI loss
  + Note: Corresponds to an upper bound or re-aligning missing historical CSI information
* Scenario B: UCI loss, known at NW and unknown at UE, with mitigation at NW
  + Note: Corresponds to implementation-based mitigation at NW but no signaling to UE.
* Scenario C: UCI loss, known at NW and UE, with mitigation at NW and UE
  + Note: Corresponds to reset of historical CSI information at both UE and NW or any other mitigation approach enabled by signaling.
* UCI loss modeling
  + Option 1: 10% UCI loss probability on all UCI reports
  + Option 2: No UCI loss for the first UCI report of each observation window, and 10% UCI loss probability for the subsequent reports of each observation window.
  + Other values for UCI loss probability are not precluded.
* Note: The same UCI loss modeling shall be applied to the benchmark for fair comparison.

|  |  |
| --- | --- |
| *Support / Can accept* |  |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Support |
| Xiaomi | Support |
| Huawei, HiSilicon | Support in general. One clarification question to “No UCI loss for the first UCI report of each observation window” – why to skip the first UCI report for UCI loss modelling? Does it mean the observation window is deferred in case the first UCI report is missed? |
| Lenovo | Support in general, Further clarifications on Option-2 for UCI Loss modelling is helpful. |
| NEC | Support |
| Panasonic | Support |
| Futurewei | For UCI loss modelling, we are ok with Option 1. We think Option 2 is not close to real deployment scenario as UE may not know UCI loss. |
| Samsung | We are fine with the above modelling when the entire UCI is lost. However, we believe it is also important to consider partial UCI loss due to CSI dropping under Scenario C. The partial dropping could be modelled based on priority of the part of CSI, e.g., X% , Y% , Z%, 10% for dropping probability for Group 0, Group 1, Group 2 of Part II CSI and Part I CSI, respectively, in Rel-16 eType II and its equivalent payload size in AI/ML CSI and X>Y>Z>10%. |
| LG | Support |
| OPPO | Support. |
| ZTE | We are also not clear about why to skip the first UCI report for UCI loss modelling? In practical, this situation may exist, so we need to further study this situation. |
| Ericsson | Support. UCI loss modelling option 2 requires further clarification. |

### Evaluation results Case 1

Observation 111a: SGCS performance Case 1

For the evaluation of temporal domain aspects **Case 1** of AI/ML based CSI compression compared to the non-AI/ML *benchmark in terms of SGCS*,

* For Layer 1,
  + 1 source [OPPO] observes performance gain of 4.9% at CSI payload X (small payload);
  + 1 source [CMCC] observes performance gain of 29.94% at CSI payload Y (medium payload).
  + Performance gain at CSI payload Z (large payload) is TBD

For the evaluation of temporal domain aspects **Case 1** of AI/ML based CSI compression compared to the CSI compression Case 0 *in terms of SGCS*,

* For Layer 1,
  + 2 sources [Futurewei, OPPO] observe performance gain of -3.2% to 12.1% at CSI payload X (small payload)
  + 2 sources [Futurewei, OPPO] observe performance gain of 6.95-12% at CSI payload Y (medium payload)
  + Performance gain at CSI payload Z (large payload) is TBD

|  |  |
| --- | --- |
| *Company* | *Comments* |
|  |  |
|  |  |

Observation 112a: FTP traffic performance Case 1

For the evaluation of temporal domain aspects **Case 1** of AI/ML based CSI compression compared to the non-AI/ML *benchmark in terms of mean UPT under FTP traffic*, till RAN1 #118,

* For Max Rank 1, performance gains are TBD
* For Max Rank 2,
  + For RU <=39%, performance gains are TBD
  + For RU of 40-69%, performance gains are TBD.
  + For RU >= 70%, 1 source [Futurewei] observes performance gains of 26%
    - 1 source [Futurewei] observes performance gains of 26% at CSI feedback overhead A (small overhead)
    - TBD performance gains at CSI feedback overhead B (medium overhead)
    - TBD performance gains at CSI feedback overhead C (large overhead)
* For Max Rank 4, performance gains are TBD.

For the evaluation of temporal domain aspects **Case 1** of AI/ML based CSI compression compared to the Case 0 *benchmark in terms of mean UPT under FTP traffic*, till RAN1 #118,

* For Max Rank 1, performance gains are TBD
* For Max Rank 2,
  + For RU <=39%, performance gains are TBD
  + For RU of 40-69%, performance gains are TBD.
  + For RU >= 70%, 1 source [Futurewei] observes performance gains of 2%
    - 1 source [Futurewei] observes performance gains of 2% at CSI feedback overhead A (small overhead)
    - TBD performance gains at CSI feedback overhead B (medium overhead)
    - TBD performance gains at CSI feedback overhead C (large overhead)
* For Max Rank 4, performance gains are TBD.

For the evaluation of temporal domain aspects **Case 1** of AI/ML based CSI compression compared to the non-AI/ML *benchmark in terms of 5% UPT under FTP traffic*, till RAN1 #118,

* For Max Rank 1, performance gains are TBD
* For Max Rank 2,
  + For RU <=39%, performance gains are TBD
  + For RU of 40-69%, performance gains are TBD.
  + For RU >= 70%, 1 source [Futurewei] observes performance gains of 73%
    - 1 source [Futurewei] observes performance gains of 73% at CSI feedback overhead A (small overhead)
    - TBD performance gains at CSI feedback overhead B (medium overhead)
    - TBD performance gains at CSI feedback overhead C (large overhead)
* For Max Rank 4, performance gains are TBD.

For the evaluation of temporal domain aspects **Case 1** of AI/ML based CSI compression compared to the Case 0 *benchmark in terms of 5% UPT under FTP traffic*, till RAN1 #118,

* For Max Rank 1, performance gains are TBD
* For Max Rank 2,
  + For RU <=39%, performance gains are TBD
  + For RU of 40-69%, performance gains are TBD.
  + For RU >= 70%, 1 source [Futurewei] observes performance gains of 12%
    - 1 source [Futurewei] observes performance gains of 12% at CSI feedback overhead A (small overhead)
    - TBD performance gains at CSI feedback overhead B (medium overhead)
    - TBD performance gains at CSI feedback overhead C (large overhead)
* For Max Rank 4, performance gains are TBD.

|  |  |
| --- | --- |
| *Company* | *Comments* |
|  |  |
|  |  |

Observation 113a: Full buffer Case 1

|  |  |
| --- | --- |
| *Company* | *Comments* |
|  |  |
|  |  |

Observation 114a: CSI feedback reduction Case 1

For the evaluation of temporal domain aspects **Case 2** of AI/ML based CSI compression, compared to the non-AI/ML benchmark, in terms of CSI feedback reduction, till RAN1 #118,

* For Max rank = 1, CSI feedback reduction is TBD.
* For Max rank = 2,
  + For CSI feedback overhead A (small overhead), 1 source [Futurewei] observes CSI feedback reduction of 92%
* For Max rank = 4, CSI feedback reduction is TBD.

For the evaluation of temporal domain aspects **Case 2** of AI/ML based CSI compression, compared to the Case 0 benchmark, in terms of CSI feedback reduction, till RAN1 #118,

* CSI feedback reduction is TBD.

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

* Precoding matrix is used as the model input.
* Training data samples are not quantized, i.e., Float32 is used/represented.
* 1-on-1 joint training is assumed.
* The performance metric is SGCS for Layer 1 of Max rank 1 or Layer 1/2 of Max rank 2.
* CSI payload X is ≤ 80/α bits; CSI payload Y is (100 - 140 )/α bits; CSI payload Z is ≥ 230/α bits; where X, Y, Z are applicable per layer, where alpha = 1 for rank = 1/2 and alpha = 2 for rank = 3/4.
* Benchmark is Rel-16 Type II codebook.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | Is it a typo? Case 2=>Case 1?  For the evaluation of temporal domain aspects **Case 1~~2~~** of AI/ML based CSI compression, compared to the non-AI/ML benchmark, in terms of CSI feedback reduction, till RAN1 #118,  …  For the evaluation of temporal domain aspects **Case 1~~2~~** of AI/ML based CSI compression, compared to the Case 0 benchmark, in terms of CSI feedback reduction, till RAN1 #118, |
|  |  |

### Evaluation results Case 2

Observation 121a: SGCS performance Case 2

For the evaluation of temporal domain aspects Case 2 of AI/ML based CSI compression compared to the non-AI/ML *benchmark in terms of SGCS*,

For Layer 1,

* 13 sources [Fujitsu, ZTE, Apple, QC, Samsung, vivo, OPPO, Xiaomi, Spreadtrum, Huawei, ETRI, Nokia, Futurewei] observe performance gain of 9.12-27.8% at CSI payload X (small payload), for which the median SGCS gain is 14.8%;
* 10 sources [ZTE, QC, vivo, CATT, CMCC, Xiaomi, Spreadtrum, Huawei, ETRI, Nokia] observe performance gain of 4.34-27.9% at CSI payload Y (medium payload), for which the median SGCS gain is 11%;
* 8 sources [ZTE, QC, vivo, Huawei, CATT, Xiaomi, Spreadtrum, Nokia] observes performance gain of 3.4-17.2% at CSI payload Z (large payload), for which the median SGCS gain is 6.26%.
* The following boxchart shows the median, 0.75 quantile, 0.25 quantile, outliers, and min/max values excluding outliers, for (X,Y,Z) CSI payload bins.

A graph with blue squares and white text

Description automatically generated

For Layer 2,

* 8 sources [ZTE, QC, Samsung, vivo, Huawei, Xiaomi, Nokia, Futurewei] observe performance gain between 14.2-37.5% at CSI payload X (small payload) , for which the median SGCS gain is 19.67%.
* 6 sources [ZTE, QC, vivo, Huawei, Nokia, Xiaomi] observe performance gain of 3.22-30% at CSI payload Y (medium payload), for which the median SGCS gain is 17%.
* 6 sources [ZTE, QC, vivo, Huawei, Nokia, Xiaomi] observe performance gain of 5.17-18% at CSI payload Z (large payload) , for which the median SGCS gain is 13.07%.
* The following boxchart shows the median, 0.75 quantile, 0.25 quantile, outliers, and min/max values excluding outliers, for (X,Y,Z) CSI payload bins.

A graph with blue squares and white text

Description automatically generated

For Layer 3,

* 2 sources [QC, Samsung] observe performance gain between 29.95-146.8% at CSI payload X (small payload)
* 1 source [QC] observes performance gain of 44.9% at CSI payload Y (medium payload)
* 1 source [QC] observes performance gain of 23.7% at CSI payload Z (large payload)

For Layer 4,

* 2 sources [QC, Samsung] observe performance gain between 38.55-280.6% at CSI payload X (small payload)
* 1 source [QC] observes performance gain of 59% at CSI payload Y (medium payload)
* 1 source [QC] observes performance gain of 33.5% at CSI payload Z (large payload)

For the evaluation of temporal domain aspects **Case 2** of AI/ML based CSI compression compared to the CSI compression Case 0 *in terms of SGCS*,

For Layer 1,

* 14 sources [Fujitsu, ZTE, Apple, QC, Samsung, vivo, OPPO, Xiaomi, Spreadtrum, Huawei, ETRI, Nokia, Futurewei, InterDigital] observe performance gain of 0.62-30% at CSI payload X (small payload) , for which the median SGCS gain is 8.6%.
* 11 sources [ZTE, QC, vivo, CATT, CMCC, Xiaomi, Spreadtrum, Huawei, ETRI, Nokia, IIT Kanpur] observe performance gain of 1.49-36% at CSI payload Y (medium payload) , for which the median SGCS gain is 6.02%.
* 9 sources [ZTE, QC, vivo, Huawei, CATT, Xiaomi, Spreadtrum, Nokia, IIT Kanpur] observe performance gain of 0.5-6.78% at CSI payload Z (large payload) , for which the median SGCS gain is 4.3%.
* The following boxchart shows the median, 0.75 quantile, 0.25 quantile, outliers, and min/max values excluding outliers, for (X, Y, Z) CSI payload bins.

A graph with blue squares and white text

Description automatically generated

For Layer 2,

* 9 sources [ZTE, QC, Samsung, vivo, Huawei, Xiaomi, Nokia, Futurewei, InterDigital] observe performance gain of 0.61-20% at CSI payload X (small payload), for which the median SGCS gain is 9.73%.
* 6 sources [ZTE, QC, vivo, Huawei, Nokia, Xiaomi] observe performance gain of 1.06-16.49% at CSI payload Y (medium payload) , for which the median SGCS gain is 10%.
* 6 sources [ZTE, QC, vivo, Huawei, Nokia, Xiaomi] observe performance gain of 2.98-13% at CSI payload Z (large payload) , for which the median SGCS gain is 7.95%.
* The following boxchart shows the median, 0.75 quantile, 0.25 quantile, outliers, and min/max values excluding outliers, for (X, Y, Z) CSI payload bins.

A graph of a number of points

Description automatically generated with medium confidence

For Layer 3,

* 2 source [QC, Samsung] observe performance gain between 8.7-29.96% at CSI payload X (small payload)
* 1 source [QC] observes performance gain of 8.4% at CSI payload Y (medium payload)
* 1 source [QC] observes performance gain of 7.9% at CSI payload Z (large payload)

For Layer 4,

* 2 source [QC, Samsung] observes performance gain between 8.33-10.9% at CSI payload X (small payload)
* 1 source [QC] observes performance gain of 8.7% at CSI payload Y (medium payload)
* 1 source [QC] observes performance gain of 7.6% at CSI payload Z (large payload)

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

* Precoding matrix (SVD output or in angle-delay domain) is used as the model input.
* Training data samples are not quantized, i.e., Float32 is used/represented.
* 1-on-1 joint training is assumed.
* The performance metric is SGCS for Layer 1 of Max rank 1, Layer 1/2 of Max rank 2, Layer 1/2/3/4 of Max Rank 4.
* CSI payload X is ≤ 80/α bits; CSI payload Y is (100 - 140 )/α bits; CSI payload Z is ≥ 230/α bits; where X, Y, Z are applicable per layer, where alpha = 1 for rank = 1/2 and alpha = 2 for rank = 3/4.
* Benchmark is Rel-16 Type II codebook.

The description of the above boxcharts is as follows:

* The line inside of each box is the median
* The top and bottom box edges represent the 0.75 quantile (upper quartile) and 0.25 quantile (lower quartile), respectively.
* Interquartile range (IQR) is the distance between the top and bottom box edges, which is used to determine the outliers (values more than 1.5 IQR away from the box edge).
* Outliers are represented by the ‘o’ marks.
* The whiskers represents the minimum and maximum values excluding outliers

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | There two benchmarks mentioned in the main text, so the R16 benchmark should be removed from the note part.  The above results are based on the following assumptions besides the assumptions of the agreed EVM table:   * Precoding matrix (SVD output or in angle-delay domain) is used as the model input. * Training data samples are not quantized, i.e., Float32 is used/represented. * 1-on-1 joint training is assumed. * The performance metric is SGCS for Layer 1 of Max rank 1, Layer 1/2 of Max rank 2, Layer 1/2/3/4 of Max Rank 4. * CSI payload X is ≤ 80/α bits; CSI payload Y is (100 - 140 )/α bits; CSI payload Z is ≥ 230/α bits; where X, Y, Z are applicable per layer, where alpha = 1 for rank = 1/2 and alpha = 2 for rank = 3/4. * ~~Benchmark is Rel-16 Type II codebook.~~ |
| ZTE | For the first bullet in the assumptions “Precoding matrix (SVD output or in angle-delay domain) is used as the model input.” To our understanding, only one company provides the GCS performance in angular-delay domain precoding matrix, while the majority of companies use SVD output as model input. So, I think it is better to add a Note to capture the results for angular-delay domain input independently, which is a different approach from other companies, and suggest not incorporated with the results for SVD input. |

Observation 122a: FTP traffic performance Case 2

For the evaluation of temporal domain aspects **Case 2** of AI/ML based CSI compression compared to the non-AI/ML *benchmark in terms of mean UPT* *under FTP* traffic, till RAN1 #118,

For Max Rank 1,

* For RU <= 39%, 2 sources [Huawei, QC] observes performance gain of 0-3.4%:
  + 2 sources [Huawei, QC] observes performance gain of 0-3.4% at CSI feedback overhead A (small overhead)
  + 2 sources [Huawei, QC] observes performance gain of 0-2.4% at CSI feedback overhead B (medium overhead)
  + 2 source [Huawei, QC] observes performance gain if 0-2% at CSI feedback overhead C (large overhead)
* For RU 40-69%, 3 sources [Huawei, Spreadtrum, QC] observed performance gain of 2-6%
  + 3 sources [Huawei, Spreadtrum, QC] observe performance gain of 3-6% at CSI feedback overhead A (small overhead)
  + 3 sources [Huawei, Spreadtrum, QC] observe performance gain of 2-5% at CSI feedback overhead B (medium overhead)
  + 3 sources [Huawei, Spreadtrum, QC] observe performance gain of 2-6% at CSI feedback overhead C (large overhead)
* For RU > 70%, 4 sources [Huawei, Spreadtrum, QC, Oppo] observes performance gain of 3-15%
  + 4 sources [Huawei, Spreadtrum, QC, Oppo] observed performance gain of 3-15% at CSI feedback overhead A (small overhead)
  + 4 sources [Huawei, Spreadtrum, QC, Oppo] observes performance gain of 4-8% at CSI feedback overhead B (medium overhead)
  + 4 sources [Huawei, Spreadtrum, QC, Oppo] observes performance gain of 4-8% at CSI feedback overhead C (large overhead)

For Max Rank 2,

* For RU <= 39%, 5 sources [Huawei, Interdigital, QC, Nokia, ZTE] observes performance gain of 0-12%:
  + 5 sources [Huawei, Interdigital, QC, Nokia, ZTE] observe performance gain of 1-12% at CSI feedback overhead A (small overhead)
  + 4 sources [Huawei, QC, Nokia, ZTE] observes performance gain of 0-3.3% at CSI feedback overhead B (medium overhead)
  + 4 source [Huawei, QC, Nokia, ZTE] observes performance gain if 0-5.6% at CSI feedback overhead C (large overhead)
* For RU 40-69%, 5 sources [Huawei, Interdigital, QC, Nokia, ZTE] observed performance gain of 1-17%
  + 5 sources [Huawei, Interdigital, QC, Nokia, ZTE] observe performance gain of 4-17% at CSI feedback overhead A (small overhead)
  + 4 sources [Huawei, QC, Nokia, ZTE] observes performance gain of 2-11% at CSI feedback overhead B (medium overhead)
  + 4 sources [Huawei, QC, Nokia, ZTE] observes performance gain of 1-9% at CSI feedback overhead C (large overhead)
* For RU > 70%, 6 sources [Huawei, Interdigital, Futurewei, QC, Nokia, ZTE] observe performance gain of 3-29%
  + 6 sources [Huawei, Interdigital, Futurewei, QC, Nokia, ZTE] sources observed performance gain of 6-29% at CSI feedback overhead A (small overhead)
  + 4 sources [Huawei, QC, Nokia, ZTE] observes performance gain of 3-17% at CSI feedback overhead B (medium overhead)
  + 4 source [Huawei, QC, Nokia, ZTE] observes performance gain of 3-17% at CSI feedback overhead C (large overhead)

For Max Rank 4,

* For RU <= 39%, 1 source [QC] observes a performance gain of 0-6%
  + 1 source [QC] observes a performance gain of 5.5% at CSI feedback overhead A (small overhead).
  + 1 source [QC] observes a performance gain of 5.3% at CSI feedback overhead B (medium overhead).
  + 1 source [QC] observes a performance gain of 0.1% at CSI feedback overhead C (large overhead).
* For RU 40-69%, 1 source [QC] observes a performance gain of 8-18%
  + 1 source [QC] observes a performance gain of 17.5% at CSI feedback overhead A (small overhead).
  + 1 source [QC] observes a performance gain of 16.4% at CSI feedback overhead B (medium overhead).
  + 1 source [QC] observes a performance gain of 8.8% at CSI feedback overhead C (large overhead).
* For RU > 70%, 1 source [QC] observes a performance gain of 12-24%
  + 1 source [QC] observes a performance gain of 23.5% at CSI feedback overhead A (small overhead).
  + 1 source [QC] observes a performance gain of 22.7% at CSI feedback overhead B (medium overhead).
  + 1 source [QC] observes a performance gain of 12.9% at CSI feedback overhead C (large overhead).

For the evaluation of temporal domain aspects **Case 2** of AI/ML based CSI compression compared to the CSI compression Case 0 *~~benchmark~~ in terms of mean UPT* *under FTP* traffic, till RAN1 #118,

For Max Rank 1,

* For RU <= 39%, 1 source [Huawei] observes performance gain of 0.8-1.6%:
  + 1 source [Huawei] observes performance gain of 1.6% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes performance gain of 1.4% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes performance gain if 0.8% at CSI feedback overhead C (large overhead)
* For RU 40-69%, 3 sources [Huawei, Spreadtrum, QC] observed performance gain of 0-3%
  + 3 sources [Huawei, Spreadtrum, QC] observes performance gain of 0-3% at CSI feedback overhead A (small overhead)
  + 3 sources [Huawei, Spreadtrum, QC] observes performance gain of 0-2% at CSI feedback overhead B (medium overhead)
  + 3 sources [Huawei, Spreadtrum, QC] observes performance gain of 0-1.4% at CSI feedback overhead C (large overhead)
* For RU > 70%, 4 sources [Huawei, Spreadtrum, QC, Oppo] observes performance gain of 0-6%
  + 4 sources [Huawei, Spreadtrum, QC, Oppo] observed performance gain of 0-6% at CSI feedback overhead A (small overhead)
  + 3 sources [Huawei, Spreadtrum, QC] observes performance gain of 2-3.3% at CSI feedback overhead B (medium overhead)
  + 3 sources [Huawei, Spreadtrum, QC] observes performance gain of 0.9-1.4% at CSI feedback overhead C (large overhead)

For Max Rank 2,

* For RU <= 39%, 5 sources [Huawei, Interdigital, QC, Nokia, ZTE] observe performance gain of -1% to 3.3%:
  + 4 sources[Huawei, Interdigital, Nokia, ZTE] observe performance gain of -1% to 3.3% at CSI feedback overhead A (small overhead)
  + 3 source [Huawei, Nokia, ZTE] observes performance gain of 0.17%-2% at CSI feedback overhead B (medium overhead)
  + 3 sources [Huawei, Nokia, ZTE] observes performance gain if -0.05% to 2% at CSI feedback overhead C (large overhead)
* For RU 40-69%, 6 sources [Huawei, Interdigital, QC, Nokia, Futurewei, ZTE] observed performance gain of -2% to 7%
  + 6 sources [Huawei, Interdigital, QC, Nokia, Futurewei, ZTE] observe performance gain of -2% to 7% at CSI feedback overhead A (small overhead)
  + 4 source [Huawei, QC, Nokia, ZTE] observes performance gain of 0-4.2% at CSI feedback overhead B (medium overhead)
  + 4 source [Huawei, QC, Nokia, ZTE] observes performance gain of 0-3% at CSI feedback overhead C (large overhead)
* For RU > 70%, 6 sources [Huawei, Futurewei, Interdigital, QC, Nokia, ZTE] observes performance gain of -5% to 14%
  + 6 sources Huawei, Futurewei, Interdigital, QC, Nokia, ZTE] observed performance gain of -5% to 14% at CSI feedback overhead A (small overhead)
  + 4 sources [Huawei, QC, Nokia, ZTE] observes performance gain of 1-12% at CSI feedback overhead B (medium overhead)
  + 4 sources [Huawei, QC, Nokia, ZTE] observes performance gain of 0-10% at CSI feedback overhead C (large overhead)

For Max Rank 4,

* For RU <= 39%, 1 source [QC] observes performance gain of -0.2% to 0.9%:
  + 1 source [QC] observes a performance gain of -0.2% at CSI feedback overhead A (small overhead).
  + 1 source [QC] observes a performance gain of 0.5% at CSI feedback overhead B (medium overhead).
  + 1 source [QC] observes a performance gain of 0.9% at CSI feedback overhead C (large overhead).
* For RU 40-69%, 1 source [QC] observes a performance gain of 4.2-6.5%
  + 1 source [QC] observes a performance gain of 6.5% at CSI feedback overhead A (small overhead).
  + 1 source [QC] observes a performance gain of 6.0% at CSI feedback overhead B (medium overhead).
  + 1 source [QC] observes a performance gain of 4.2% at CSI feedback overhead C (large overhead).
* For RU > 70%, 1 source [QC] observes a performance gain of 7.5-11.1%
  + 1 source [QC] observes a performance gain of 11.1% at CSI feedback overhead A (small overhead).
  + 1 source [QC] observes a performance gain of 10.9% at CSI feedback overhead B (medium overhead).
  + 1 source [QC] observes a performance gain of 7.5% at CSI feedback overhead C (large overhead)..

For the evaluation of temporal domain aspects **Case 2** of AI/ML based CSI compression compared to the non-AI/ML *benchmark in terms of 5% UPT under FTP*, till RAN1 #118,

For Max rank 1:

* For RU <= 39%, 2 sources [Huawei, QC] observes performance gain of 0-8%:
  + 2 sources [Huawei, QC] observe the performance gain of 1.5-8% at CSI feedback overhead A (small overhead)
  + 2 sources [Huawei, QC] observes the performance gain of 0.3%-4% at CSI feedback overhead B (medium overhead)
  + 2 sources [Huawei, QC] observes the performance gain of 0.7-4% at CSI feedback overhead C (large overhead)
* For RU between 40-69%, 3 sources [Huawei, Spreadtrum, QC] observes performance gain of 4-12%:
  + 3 sources [Huawei, Spreadtrum, QC] observes the performance gain of 10-12% at CSI feedback overhead A (small overhead)
  + 3 sources [Huawei, Spreadtrum, QC] observes the performance gain of ~8% at CSI feedback overhead B (medium overhead)
  + 3 source [Huawei, Spreadtrum, QC] observes the performance gain of 4-7% at CSI feedback overhead C (large overhead)
* For RU > 70%, 3 sources [Huawei, Spreadtrum, QC] observes performance gain of 10-37%:
  + 3 sources [Huawei, Spreadtrum, QC] observes the performance gain of 10-37% at CSI feedback overhead A (small overhead)
  + 3 sources [Huawei, Spreadtrum, QC] observes the performance gain of 12-27% at CSI feedback overhead B (medium overhead)
  + 3 sources [Huawei, Spreadtrum, QC] observes the performance gain of 10-30% at CSI feedback overhead C (large overhead)

For Max Rank 2:

* For RU <= 39%, 5 sources [Huawei, Interdigital, QC, Nokia, ZTE] observe performance gain of 1-45%:
  + 5 sources [Huawei, Interdigital, QC, Nokia, ZTE] observe the performance gain of 6-45% at CSI feedback overhead A (small overhead)
  + 4 sources [Huawei, QC, Nokia, ZTE] observes the performance gain of 1.5-14% at CSI feedback overhead B (medium overhead)
  + 4 sources [Huawei, QC, Nokia, ZTE] observes the performance gain of 1-8% at CSI feedback overhead C (large overhead)
* For RU between 40-69%, 5 sources [Huawei, Interdigital, QC, Nokia, ZTE] observe performance gain of 3-41%:
  + 5 sources [Huawei, Interdigital, QC, Nokia, ZTE] observe the performance gain of 9-41% at CSI feedback overhead A (small overhead)
  + 4 sources [Huawei, QC, Nokia, ZTE] observes the performance gain of 3-33% at CSI feedback overhead B (medium overhead)
  + 4 sources [Huawei, QC, Nokia, ZTE] observes the performance gain of 3-21% at CSI feedback overhead C (large overhead)
* For RU > 70%, 6 sources [Huawei, Futurewei, Interdigital, QC, Nokia, ZTE] observe performance gain of 6-73%:
  + 6 sources [Huawei, Futurewei, Interdigital, QC, Nokia, ZTE] observe the performance gain of 14-70% at CSI feedback overhead A (small overhead)
  + 4 sources [Huawei, QC, Nokia, ZTE] observes the performance gain of 5-51% at CSI feedback overhead B (medium overhead)
  + 4 source [Huawei, QC, Nokia, ZTE] observes the performance gain of 6-32% at CSI feedback overhead C (large overhead)

For Max Rank 4:

* For RU <= 39%, 1 source [QC] observes performance gain of 10-23%:
  + 1 source [QC] observes a performance gain of 22.5% at CSI feedback overhead A (small overhead).
  + 1 source [QC] observes a performance gain of 18.9% at CSI feedback overhead B (medium overhead).
  + 1 source [QC] observes a performance gain of 10.2% at CSI feedback overhead C (large overhead).
* For RU 40-69%, 1 source [QC] observes a performance gain of 33-56%
  + 1 source [QC] observes a performance gain of 55.4% at CSI feedback overhead A (small overhead).
  + 1 source [QC] observes a performance gain of 48.1% at CSI feedback overhead B (medium overhead).
  + 1 source [QC] observes a performance gain of 33.1% at CSI feedback overhead C (large overhead).
* For RU > 70%, 1 source [QC] observes a performance gain of 47-79%
  + 1 source [QC] observes a performance gain of 79% at CSI feedback overhead A (small overhead).
  + 1 source [QC] observes a performance gain of 69.9% at CSI feedback overhead B (medium overhead).
  + 1 source [QC] observes a performance gain of 47.2% at CSI feedback overhead C (large overhead)..

For the evaluation of temporal domain aspects **Case 2** of AI/ML based CSI compression compared to the CSI compression Case 0 *~~benchmark~~ in terms of 5% UPT under FTP*, till RAN1 #118,

For Max rank 1,

* For RU <= 39%, 1 source [Huawei] observes performance gain of 1-5%:
  + 1 source [Huawei] observes the performance gain of 5% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes the performance gain of 3% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes the performance gain of 1% at CSI feedback overhead C (large overhead)
* For RU between 40-69%, 3 sources [Huawei, Spreadtrum, QC] observes performance gain of 1-5%:
  + 3 sources [Huawei, Spreadtrum, QC] observes the performance gain of 3.6-5% at CSI feedback overhead A (small overhead)
  + 3 sources [Huawei, Spreadtrum, QC] observes the performance gain of 2.8-5% at CSI feedback overhead B (medium overhead)
  + 3 sources [Huawei, Spreadtrum, QC] observes the performance gain of 1.2-4% at CSI feedback overhead C (large overhead)
* For RU > 70%, 1 source [Huawei] observes performance gain of 1-13%:
  + 3 sources [Huawei, Spreadtrum, QC] observe the performance gain of 7-10% at CSI feedback overhead A (small overhead)
  + 3 sources [Huawei, Spreadtrum, QC] observe the performance gain of 4-12.9% at CSI feedback overhead B (medium overhead)
  + 3 sources [Huawei, Spreadtrum, QC] observe the performance gain of 1-6.5% at CSI feedback overhead C (large overhead)

For Max Rank 2,

* For RU <= 39%, 4 sources [Huawei, Interdigital, Nokia, ZTE] observes performance gain of 0-12%:
  + 4 sources [Huawei, Interdigital, Nokia, ZTE] observes the performance gain of 2-12% at CSI feedback overhead A (small overhead)
  + 3 sources [Huawei, Nokia, ZTE] observes the performance gain of 0-5% at CSI feedback overhead B (medium overhead)
  + 3 source [Huawei, Nokia, ZTE] observes the performance gain of 1-3% at CSI feedback overhead C (large overhead)
* For RU between 40-69%, 5 sources [Huawei, Interdigital, QC, Nokia, ZTE] observes performance gain of 0-17%:
  + 5 sources [Huawei, Interdigital, QC, Nokia, ZTE] observes the performance gain of 4-17% at CSI feedback overhead A (small overhead)
  + 4 sources [Huawei, QC, Nokia, ZTE] observes the performance gain of 1-14% at CSI feedback overhead B (medium overhead)
  + 4 sources [Huawei, QC, Nokia, ZTE] observes the performance gain of 0-13% at CSI feedback overhead C (large overhead)
* For RU > 70%, 6 sources [Huawei, Futurewei, Interdigital, QC, Nokia, ZTE] observes performance gain of 3-30%:
  + 6 sources [Huawei, Futurewei, Interdigital, QC, Nokia, ZTE] observes the performance gain of 8-30% at CSI feedback overhead A (small overhead)
  + 4 sources [Huawei, QC, Nokia, ZTE] observes the performance gain of 3-22% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei, QC, Nokia, ZTE] observes the performance gain of 2-17% at CSI feedback overhead C (large overhead)

For Max Rank 4:

* For RU <= 39%, 1 source [QC] observes performance gain of 4-10.1%:
  + 1 source [QC] observes a performance gain of 4.6% at CSI feedback overhead A (small overhead).
  + 1 source [QC] observes a performance gain of 10.1% at CSI feedback overhead B (medium overhead).
  + 1 source [QC] observes a performance gain of 6.8% at CSI feedback overhead C (large overhead).
* For RU 40-69%, 1 source [QC] observes a performance gain of 4-17%
  + 1 source [QC] observes a performance gain of 16.4% at CSI feedback overhead A (small overhead).
  + 1 source [QC] observes a performance gain of 15% at CSI feedback overhead B (medium overhead).
  + 1 source [QC] observes a performance gain of 4.6% at CSI feedback overhead C (large overhead).
* For RU > 70%, 1 source [QC] observes a performance gain of 11-19.1%
  + 1 source [QC] observes a performance gain of 14.3% at CSI feedback overhead A (small overhead).
  + 1 source [QC] observes a performance gain of 19.1% at CSI feedback overhead B (medium overhead).
  + 1 source [QC] observes a performance gain of 11.1% at CSI feedback overhead C (large overhead)..

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | Add the description to be aligned with other observations?  The above results are based on the following assumptions besides the assumptions of the agreed EVM table:  - Precoding matrix (SVD output or in angle-delay domain) is used as the model input.  - Training data samples are not quantized, i.e., Float32 is used/represented.  - 1-on-1 joint training is assumed.  - The performance metric is UPT for Max rank 1, Max rank 2, or Max rank 4. |
| ZTE | We think the median value for UPT can be also calculated to keep aligned with SGCS statistics if the number of companies is larger than 2.  In addition, a typo in the 3rd bullet of “Max rank 2, RU>70%” should be modified as  ~~1~~4 sources [Huawei, QC, Nokia, ZTE] observes the performance gain of 2-17% at CSI feedback overhead C (large overhead) |

Observation 123a: Full buffer performance Case 2

For the evaluation of temporal domain aspects **Case 2** of AI/ML based CSI compression compared to the non-AI/ML *benchmark, in terms of mean UPT under full buffer*, till RAN1 #118,

* For Max Rank 1, 5 sources [Huawei, QC, Oppo, Xiaomi and Vivo] observe performance gains of 0-25%
  + 5 sources [Huawei, QC, Oppo, Xiaomi and Vivo] observe performance gains of 0-25% at CSI feedback overhead A (small overhead)
  + 5 source [Huawei, QC, Oppo, Xiaomi and Vivo] observes performance gains of 0-20% at CSI feedback overhead B (medium overhead)
  + 5 source [Huawei, QC, Oppo, Xiaomi and Vivo] observes performance gains of 0-18% at CSI feedback overhead C (large overhead)
* For Max Rank 2, 7 sources [Huawei, Fujitsu, Xiaomi, QC, Vivo, Nokia, ZTE] observe performance gains of 1-30%
  + 7 sources [Huawei, Fujitsu, Xiaomi, QC, Vivo, Nokia, ZTE] observe performance gains of 6-30% at CSI feedback overhead A (small overhead)
  + 7 sources [Huawei, Fujitsu, Xiaomi, QC, Vivo, Nokia, ZTE] observe performance gains of 3-23% at CSI feedback overhead B (medium overhead)
  + 7 sources [Huawei, Fujitsu, Xiaomi, QC, Vivo, Nokia, ZTE] observe performance gains of 2-24% at CSI feedback overhead C (large overhead)

The following boxchart shows the median, 0.75 quantile, 0.25 quantile, outliers, and min/max values excluding outliers, for (A, B, C) CSI payload bins. The source, ZTE, has 2 submissions, hence the number of points is higher than 7 in some cases.



Figure 1: Mean Throughput gains over Benchmark 1, for Bins A, B and C, for Max Rank = 2.

* For Max Rank 4, 2 sources [QC, Samsung] observe performance gains of 9-16%
  + 2 sources [QC, Samsung] observe performance gains of 9-16% at CSI overhead A (small overhead)
  + 1 source [QC] observes performance gains of 14.3% at CSI overhead B (medium overhead)
  + 1 source [QC] observes performance gains of 9.2% at CSI overhead C (large overhead)

For the evaluation of temporal domain aspects **Case 2** of AI/ML based CSI compression compared to the CSI compression Case 0 *~~benchmark~~, in terms of mean UPT under full buffer*, till RAN1 #117,

* For Max Rank 1, 6 sources [Huawei, IIT Kanpur, QC, Oppo, Xiaomi and Vivo] observe performance gains of 0-13%
  + 5 sources [Huawei, QC, Oppo, Xiaomi and Vivo] observe performance gains of 0-13% at CSI feedback overhead A (small overhead)
  + 6 sources [Huawei, IIT Kanpur, QC, Oppo, Xiaomi and Vivo] observe performance gains of 0-12% at CSI feedback overhead B (medium overhead)
  + 2 sources [Huawei, IIT Kanpur, QC, Oppo, Xiaomi and Vivo] observe performance gains of 0-7.7% at CSI feedback overhead C (large overhead)
* For Max Rank 2, 7 sources [Huawei, Fujitsu, Xiaomi, QC, Vivo, Nokia, ZTE] observe performance gains of 1-14%
  + 6 sources [Huawei, Fujitsu, Xiaomi, QC, Nokia, ZTE] observe performance gains of 2-14% at CSI feedback overhead A (small overhead)
  + 5 sources [Huawei, Xiaomi, QC, Nokia, ZTE] observe performance gains of 1-14% at CSI feedback overhead B (medium overhead)
  + 5 sources [Huawei, Xiaomi, QC, Nokia, ZTE] observe performance gains of 1-9% at CSI feedback overhead C (large overhead)

The following boxchart shows the median, 0.75 quantile, 0.25 quantile, outliers, and min/max values excluding outliers, for (A, B, C) CSI payload bins.



Figure 2: Mean Throughput gains over Benchmark 2, for Bins A, B and C, for Max Rank =2.

* For Max Rank 4, 1 source [QC] observes performance gains of 8-12%
  + 1 source [QC] observes performance gains of 11.3% at CSI overhead A (small overhead)
  + 1 source [QC] observes performance gains of 14.3% at CSI overhead B (medium overhead)
  + 1 source [QC] observes performance gains of 9.2% at CSI overhead C (large overhead)

For the evaluation of temporal domain aspects **Case 2** of AI/ML based CSI compression compared to the non-AI/ML *benchmark in terms of 5% UPT under full buffer*, till RAN1 #118,

* For Max Rank 1, 2 sources [QC, vivo] observe performance gains of 7-16%
  + 2 sources [QC, vivo] observe performance gain of 10-15.3% at CSI feedback overhead A (small overhead)
  + 2 sources [QC, vivo) observe performance gain of 7.5-8.1% at CSI feedback overhead B (medium overhead)
  + 2 sources [QC, vivo] observe performance gain of 7.2-8.5% at CSI feedback overhead C (large overhead)
* For Max Rank 2, 5 sources [Fujitsu, QC, vivo, Nokia, ZTE] observes performance gains of 6-58%
  + 5 sources [Fujitsu, QC, vivo, Nokia, ZTE] observe performance gains of 12-58% at CSI feedback overhead A (small overhead)
  + 4 sources [QC, Vivo, Nokia, ZTE] observe performance gain of 8-12% at CSI feedback overhead B (medium overhead)
  + 4 sources [QC, Vivo, Nokia, ZTE] observe performance gain of 2-9% at CSI feedback overhead C (large overhead).

The following boxchart shows the median, 0.75 quantile, 0.25 quantile, outliers, and min/max values excluding outliers, for (A, B, C) CSI payload bins.



Figure 3: Edge Throughput gains over Benchmark 1, for Bins A, B and C, for Max Rank =2

* For Max Rank 4, 1 source [QC] observes performance gain of 4.3-10.2%
  + 1 source [QC] observes performance gain of 10.2% at CSI feedback overhead A (small overhead)
  + 1 source [QC] observes performance gain of 7.4% at CSI feedback overhead B (medium overhead)
  + 1 source [QC] observes performance gain of 4.3% at CSI feedback overhead C (large overhead)

For the evaluation of temporal domain aspects **Case 2** of AI/ML based CSI compression compared to the CSI compression Case 0 *~~benchmark~~ in terms of 5% UPT under full buffer*, till RAN1 #118,

* For Max Rank 1, 2 sources [IIT Kanpur, QC] observe performance gains of 2-14%.
  + 1 sources [QC] observe performance gains of 2% at CSI feedback overhead A (small overhead)
  + 2 sources [IIT Kanpur, QC] observe performance gains of 3.4-14% at CSI feedback overhead B (medium overhead)
  + 2 sources [IIT Kanpur, QC] observe performance gains of 2.8-4.54% at CSI feedback overhead C (large overhead)
* For Max Rank 2, 4 sources [QC, Fujitsu, Nokia, ZTE] observes performance gains of 2-35%
  + 4 sources [QC, Fujitsu, Nokia, ZTE] observe performance gains of 3-35% at CSI feedback overhead A (small overhead)
  + 3 sources [QC, Nokia, ZTE] observe performance gains of 2-6.31% at CSI feedback overhead B (medium overhead)
  + 3 sources [QC, Nokia, ZTE] observe performance gains of 2-8% at CSI feedback overhead C (large overhead)

The following boxchart shows the median, 0.75 quantile, 0.25 quantile, outliers, and min/max values excluding outliers, for (A, B, C) CSI payload bins.



Figure 4: Edge Throughput gains over Benchmark 2, for Bins A, B and C, for Max Rank =2

* For Max Rank 4, 1 source [QC] observes performance gain of 3-3.5%
  + 1 source [QC] observes performance gain of 3.5% at CSI feedback overhead A (small overhead)
  + 1 source [QC] observes performance gain of 3% at CSI feedback overhead B (medium overhead)
  + 1 source [QC] observes performance gain of 3% at CSI feedback overhead C (large overhead)

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | Add the description to be aligned with other observations?  The above results are based on the following assumptions besides the assumptions of the agreed EVM table:  - Precoding matrix (SVD output or in angle-delay domain) is used as the model input.  - Training data samples are not quantized, i.e., Float32 is used/represented.  - 1-on-1 joint training is assumed.  - The performance metric is throughput for Max rank 1, Max rank 2, or Max rank 4. |
| ZTE | To our understanding, the outliers can be excluded which may not be counted in the observation, to keep align with the approach to drawing observations in Rel-18. Then, the range of Whiskers is the same as the range of min/max range. |

Observation 124a: CSI feedback reduction Case 2

For the evaluation of temporal domain aspects **Case 2** of AI/ML based CSI compression, compared to the non-AI/ML benchmark, in terms of CSI feedback reduction, till RAN1 #118,

- For Max rank = 1,

- For CSI feedback overhead A (small overhead),

* 1 source [QC] observes CSI feedback reduction of 67% for full buffer;
* 1 source [QC] observes CSI feedback reduction of 17% for FTP traffic at RU <= 39%
* 1 source [QC] observes CSI feedback reduction of 73% for FTP traffic at RU of 40-69%
* 1 source [QC] observes CSI feedback reduction of 55% for FTP traffic at RU >=70%

- For CSI feedback overhead B (medium overhead),

* 1 source [QC] observes CSI feedback reduction of 67% for full buffer;
* 1 source [QC] observes CSI feedback reduction of 18% for FTP traffic at RU <= 39%
* 1 source [QC] observes CSI feedback reduction of 73% for FTP traffic at RU of 40-69%
* 1 source [QC] observes CSI feedback reduction of 55% for FTP traffic at RU >=70%

- For CSI feedback overhead C (large overhead),

* 2 sources [QC, Huawei] observes CSI feedback reduction of 68-75% for full buffer;
* 2 sources [QC, Huawei] observe the CSI feedback reduction of 47-74% for FTP traffic at RU <= 39%
* 2 sources [QC, Huawei] observes CSI feedback reduction of 78-80% for FTP traffic at RU of 40-69%
* 2 sources [QC, Huawei] observes CSI feedback reduction of 52-73% for FTP traffic at RU >=70%;

- For Max rank = 2,

- For CSI feedback overhead A (small overhead),

* 2 sources [QC, Nokia] observes CSI feedback reduction of 76-80% for full buffer;
* 1 source [QC] observes CSI feedback reduction of 38% for FTP traffic at RU <= 39%
* 1 source [QC] observes CSI feedback reduction of 83% for FTP traffic at RU of 40-69%
* 2 sources [QC, Futurewei] observe CSI feedback reduction of 69-92% at RU >= 70%

- For CSI feedback overhead B (medium overhead),

* 2 sources [QC, Nokia] observes CSI-feedback reduction of 73-80% for full buffer;
* 1 source [QC] observes CSI feedback reduction of 54% for FTP traffic at RU <= 39%
* 1 source [QC] observes CSI feedback reduction of 76% for FTP traffic at RU of 40-69%
* 1 source [QC] observe CSI feedback reduction of 67% at RU >= 70%

- For CSI feedback overhead C (large overhead),

* 3 sources [Huawei, QC, Nokia] observe the CSI feedback reduction of 70-80% for full buffer;
* 2 sources [Huawei, QC] observes the CSI feedback reduction of 5-53% for FTP traffic at RU <= 39%
* 2 sources [Huawei, QC] observes the CSI feedback reduction of 60-62% for FTP traffic at RU of 40-69%
* 2 sources [Huawei, QC] observes the CSI feedback reduction of 54-70% for FTP traffic at RU >= 70%

- For Max rank = 4,

. - For CSI feedback overhead A (small overhead),

* 1 source [QC] observes CSI feedback reduction of 70% for full buffer;
* 1 source [QC] observes CSI feedback reduction of 73% for FTP traffic at RU <= 39%
* TBD CSI feedback reduction for FTP traffic at RU of 40-69%
* 1 sources [QC] observe CSI feedback reduction of 87% at RU >= 70%

- For CSI feedback overhead B (medium overhead),

* 1 source [QC] observes CSI-feedback reduction of 68% for full buffer;
* 1 source [QC] observes CSI feedback reduction of 55% for FTP traffic at RU <= 39%
* 1 source [QC] observes CSI feedback reduction of 83% for FTP traffic at RU of 40-69%
* 1 source [QC] observe CSI feedback reduction of 80% at RU >= 70%

- For CSI feedback overhead C (large overhead),

* 1 source [QC] observe the CSI feedback reduction of 66% for full buffer;
* 1 source [QC] observes the CSI feedback reduction of 3% for FTP traffic at RU <= 39%
* 1 source [QC] observes the CSI feedback reduction of 67% for FTP traffic at RU of 40-69%
* 1 source [QC] observes the CSI feedback reduction of 69% for FTP traffic at RU >= 70%

For the evaluation of temporal domain aspects **Case 2** of AI/ML based CSI compression, compared to the CSI compression Case 0 *~~benchmark~~*, in terms of CSI feedback reduction, till RAN1 #117,

- For Max rank = 1,

- For CSI feedback overhead A (small overhead),

* 1 source [QC] observes CSI feedback reduction of 40% for full buffer;
* TBD CSI feedback reduction for FTP traffic at RU <= 39%
* 1 source [QC] observes CSI feedback reduction of 28% at RU of 40-69%
* 1 source [QC] observes CSI feedback reduction of 23% at RU >=70%

- For CSI feedback overhead B (medium overhead),

* 1 source [QC] observes CSI-feedback reduction of 51% for full buffer;
* TBD CSI feedback reduction for FTP traffic at RU <= 39%
* 1 source [QC] observes CSI feedback reduction of 31% at RU of 40-69%
* 1 source [QC] observes CSI feedback reduction of 41% at RU >=70%

- For CSI feedback overhead C (large overhead),

* 1 source [QC] observes CSI feedback reduction of 34% for full buffer;
* TBD CSI feedback reduction for FTP traffic at RU <= 39%
* 1 source [QC] observes CSI feedback reduction of 17% at RU of 40-69%
* 1 source [QC] observes CSI feedback reduction of 29% at RU >=70%;

- For Max rank = 2,

- For CSI feedback overhead A (small overhead),

* 2 sources [QC, Nokia] observe CSI feedback reduction of 45-50%;
* TBD CSI feedback reduction for FTP traffic at RU <= 39%
* 1 source [QC] observes CSI feedback reduction of 39% at RU of 40-69%
* 1 source [QC] observes CSI feedback reduction of 42% at RU >= 70%

- For CSI feedback overhead B (medium overhead),

* 2 sources [QC, Nokia] observes CSI-feedback reduction of 50-56% for full buffer;
* TBD CSI feedback reduction for FTP traffic at RU <= 39%
* 1 source [QC] observes CSI feedback reduction of 45% at RU of 40-69%
* 1 source [QC] observes CSI feedback reduction of 45% at RU >= 70%

- For CSI feedback overhead C (large overhead),

* 3 sources [Huawei, QC, Nokia] observe CSI feedback reduction of 50-60% for full buffer,
* TBD CSI feedback reduction for FTP traffic at RU <= 39%
* 1 source [QC] observes CSI feedback reduction of 39% at RU of 40-69%
* 1 source [QC] observes CSI feedback reduction of 32% at RU >= 70%;

- For Max rank = 4,

. - For CSI feedback overhead A (small overhead),

* 1 source [QC] observes CSI feedback reduction of 55% for full buffer;
* 1 source [QC] observes CSI feedback reduction of -3% for FTP traffic at RU <= 39%
* 1 source [QC] observes CSI feedback reductionof 45% for FTP traffic at RU of 40-69%
* 1 sources [QC] observe CSI feedback reduction of 50% at RU >= 70%

- For CSI feedback overhead B (medium overhead),

* 1 source [QC] observes CSI-feedback reduction of 59% for full buffer;
* 1 source [QC] observes CSI feedback reduction of 6% for FTP traffic at RU <= 39%
* 1 source [QC] observes CSI feedback reduction of 44% for FTP traffic at RU of 40-69%
* 1 source [QC] observe CSI feedback reduction of 55% at RU >= 70%

- For CSI feedback overhead C (large overhead),

* 1 source [QC] observe the CSI feedback reduction of 64% for full buffer;
* 1 source [QC] observes the CSI feedback reduction of 22% for FTP traffic at RU <= 39%
* 1 source [QC] observes the CSI feedback reduction of 49% for FTP traffic at RU of 40-69%
* 1 source [QC] observes the CSI feedback reduction of 54% for FTP traffic at RU >= 70%

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | Add the description to be aligned with other observations?  The above results are based on the following assumptions besides the assumptions of the agreed EVM table:  - Precoding matrix (SVD output or in angle-delay domain) is used as the model input.  - Training data samples are not quantized, i.e., Float32 is used/represented.  - 1-on-1 joint training is assumed.  - The performance metric is CSI feedback overhead reduction for Max rank 1, Max rank 2, or Max rank 4. |
|  |  |

### Evaluation results Case 5

Observation 151a: SGCS performance Case 5

For the evaluation of temporal domain aspects **Case 5** of AI/ML based CSI compression compared to the non-AI/ML *benchmark in terms of SGCS*,

* For Layer 1,
  + 2 sources [Fujitsu, OPPO] observe performance gain of 10.22-10.9% at CSI payload X (small payload)
  + ~~Performance gain at CSI payload Y (medium payload) is TBD~~
  + ~~Performance gain at CSI payload Z (large payload) is TBD~~

For the evaluation of temporal domain aspects **Case 5** of AI/ML based CSI compression compared to the CSI compression Case 0 *in terms of SGCS*,

* For Layer 1,
  + 2 sources [Fujitsu, OPPO] observe performance gain of 1.7-6.3% at CSI payload X (small payload)
  + 1 source [IIT Kanpur] observes performance gain of 39.5% at CSI payload Y (medium payload)
  + 1 source [IIT Kanpur] observes performance gain of 6.62% at CSI payload Z (large payload)

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

* Precoding matrix is used as the model input.
* Training data samples are not quantized, i.e., Float32 is used/represented.
* 1-on-1 joint training is assumed.
* The performance metric is SGCS for Layer 1 of Max rank 1 or Layer 1/2 of Max rank 2.
* CSI payload X is ≤ 80/α bits; CSI payload Y is (100 - 140 )/α bits; CSI payload Z is ≥ 230/α bits; where X, Y, Z are applicable per layer, where alpha = 1 for rank = 1/2 and alpha = 2 for rank = 3/4.
* Benchmark is Rel-16 Type II codebook.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | There two benchmarks mentioned in the main text, so the R16 benchmark should be removed from the note part.  The above results are based on the following assumptions besides the assumptions of the agreed EVM table:   * Precoding matrix is used as the model input. * Training data samples are not quantized, i.e., Float32 is used/represented. * 1-on-1 joint training is assumed. * The performance metric is SGCS for Layer 1 of Max rank 1 or Layer 1/2 of Max rank 2. * CSI payload X is ≤ 80/α bits; CSI payload Y is (100 - 140 )/α bits; CSI payload Z is ≥ 230/α bits; where X, Y, Z are applicable per layer, where alpha = 1 for rank = 1/2 and alpha = 2 for rank = 3/4. * ~~Benchmark is Rel-16 Type II codebook.~~ |
|  |  |

### Evaluation results Case 3

Observation 131a: SGCS performance Case 3

For the evaluation of temporal domain aspects **Case 3** of AI/ML based CSI compression compared to the *benchmark in terms of SGCS*, for the mixed scenario of 80% indoor and 20% outdoor users:

For Layer 1,

- 6 sources [oppo, vivo, QC, Fujitsu, ZTE, DOCOMO] observe the performance gain of 1.37-28% at CSI payload X (small payload), for which the median SGCS gain is 6.95%.

- 4 sources [CATT, ZTE, QC, DOCOMO] observes the performance gain of 3.9-22% at CSI payload Y (medium payload), for which the median SGCS gain is 11.05%.

- 4 sources [CATT, ZTE, QC, DOCOMO] observes the performance gain of 1.37-21% at CSI payload Z (large payload), for which the median SGCS gain is 8.2%.

- The following boxchart shows the median, 0.75 quantile, 0.25 quantile, outliers, and min/max values excluding outliers, for (X, Y, Z) CSI payload bins.

A diagram of a graph

Description automatically generated

For Layer 2,

- 4 sources [ZTE, QC, vivo, DOCOMO] observes the performance gain of 8.6-47% at CSI payload X (small payload);

- 3 sources [ZTE, QC, DOCOMO] observes the performance gain of 4.3-40% at CSI payload Y (medium payload);

- 3 sources [ZTE, QC, DOCOMO] observes the performance gain of 3.61-38% at CSI payload Z (large payload).

For Layer 3,

- 1 source [QC] observes the performance gain of 79.7% at CSI payload X (small payload);

- 1 source [QC] observes the performance gain of 28.9% at CSI payload Y (medium payload);

- 1 source [QC] observes the performance gain of 37.7% at CSI payload Z (large payload).

For Layer 4,

- 1 source [QC] observes the performance gain of 98.5% at CSI payload X (small payload);

- 1 source [QC] observes the performance gain of 33.6% at CSI payload Y (medium payload);

- 1 source [QC] observes the performance gain of 42.2% at CSI payload Z (large payload).

For the evaluation of temporal domain aspects **Case 3** of AI/ML based CSI compression compared to the *benchmark in terms of SGCS*, for the scenario of 100% outdoor users:

For Layer 1,

- 5 sources [Fujitsu, ZTE, Ericsson, Xiaomi, InterDigital] observe the performance gain of -4 to 39.76% at CSI payload X (small payload), for which the median SGCS gain is 7.44%;

- 4 sources [~~Samsung~~, ZTE, Ericsson, CMCC, Xiaomi] observe the performance gain of 1.03- 20.84% at CSI payload Y (medium payload), for which the median SGCS gain is 5.99%;

- 4 sources [~~Samsung,~~ ZTE, Ericsson, Xiaomi, InterDigital] observe the performance gain of 3.49-24.08% at CSI payload Z (large payload), for which the median SGCS gain is 5.6%.

- The following boxchart shows the median, 0.75 quantile, 0.25 quantile, outliers, and min/max values excluding outliers, for (X,Y,Z) CSI payload bins.

A graph with blue squares and white text

Description automatically generated

For Layer 2,

- 1 source [ZTE] observes the performance gain of 7.6% at CSI payload X (small payload);

- 1 source [ZTE] observes the performance gain of 6.3% at CSI payload Y (medium payload);

- 1 source [ZTE] observes the performance gain of 4.7% at CSI payload Z (large payload).

The description of the above boxcharts is as follows:

* The line inside of each box is the median
* The top and bottom box edges represent the 0.75 quantile (upper quartile) and 0.25 quantile (lower quartile), respectively.
* Interquartile range (IQR) is the distance between the top and bottom box edges, which is used to determine the outliers (values more than 1.5 IQR away from the box edge).
* Outliers are represented by the ‘o’ marks.
* The whiskers represent the minimum and maximum values excluding outliers

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | Add the description to be aligned with other observations?  The above results are based on the following assumptions besides the assumptions of the agreed EVM table:   * Precoding matrix is used as the compression model input. * Training data samples are not quantized, i.e., Float32 is used/represented. * 1-on-1 joint training is assumed. * The performance metric is SGCS for Layer 1 of Max rank 1, Layer 1/2 of Max rank 2, Layer 1/2/3/4 of Max Rank 4. * CSI payload X is ≤ 80/α bits; CSI payload Y is (100 - 140 )/α bits; CSI payload Z is ≥ 230/α bits; where X, Y, Z are applicable per layer, where alpha = 1 for rank = 1/2 and alpha = 2 for rank = 3/4. * Benchmark is Rel-18 doppler eT2. |
|  |  |

Observation 132a: FTP traffic performance Case 3

For the evaluation of temporal domain aspects **Case 3** of AI/ML based CSI compression compared to the *benchmark in terms of mean UPT* *under FTP* traffic, till RAN1 #118, for the mixed scenario of 80% indoor and 20% outdoor users:

For Max Rank 1,

* For RU <= 39%, 1 sources [CATT] observes performance gain of 0-1.3%:
  + TBD performance gain at CSI feedback overhead A (small overhead)
  + 1 source [CATT] observes a performance gain of 0.86% at CSI feedback overhead B (medium overhead)
  + 1 source [CATT] observes performance gain if 1.28% at CSI feedback overhead C (large overhead)
* For RU 40-69%, 1 sources [QC] observed performance gain of 0.6-4.5%
  + 1 sources [QC] observe performance gain of 4.5% at CSI feedback overhead A (small overhead)
  + 1 sources [QC] observe performance gain of 0.6% at CSI feedback overhead B (medium overhead)
  + 1 sources [QC] observe performance gain of 1.5% at CSI feedback overhead C (large overhead)
* For RU > 70%, 1 sources [QC] observes performance gain of 0.9-3.9%
  + 1 sources [QC] observed performance gain of 2.7% at CSI feedback overhead A (small overhead)
  + 1 sources [QC] observes performance gain of 0.9% at CSI feedback overhead B (medium overhead)
  + 1 sources [QC] observes performance gain of 3.9% at CSI feedback overhead C (large overhead)

For Max Rank 2,

* For RU <= 39%, 1 sources [NTT Docomo] observes performance gain of 4-6%:
  + 1 source [NTT Docomo] observes performance gain of 6.3% at CSI feedback overhead A (small overhead)
  + 1 source [NTT Docomo] observes a performance gain of 4.2% at CSI feedback overhead B (medium overhead)
  + 1 source [NTT Docomo] observes performance gain if 4.2% at CSI feedback overhead C (large overhead)
* For RU 40-69%, 2 sources [QC, NTT Docomo] observed performance gain of 1-12%
  + 2 sources [QC, NTT Docomo] observe performance gain of 2-12% at CSI feedback overhead A (small overhead)
  + 2 sources [QC, NTT Docomo] observe performance gain of 1-9.4% at CSI feedback overhead B (medium overhead)
  + 2 sources [QC, NTT Docomo] observe performance gain of 1.6-7.1% at CSI feedback overhead C (large overhead)
* For RU > 70%, 2 sources [QC, NTT Docomo] observes performance gain of -0.2% to 22.1%
  + 2 sources [QC, NTT Docomo] observed performance gain of -0.1% to 22.1% at CSI feedback overhead A (small overhead)
  + 2 sources [QC, NTT Docomo] observes performance gain of -0.2% to 16.1% at CSI feedback overhead B (medium overhead)
  + 2 sources [QC, NTT Docomo] observes performance gain of 4.3-15.2% at CSI feedback overhead C (large overhead)

For Max Rank 4,

* For RU <= 39%, 1 sources [QC] observes performance gain of 0.4-4.1%:
  + 1 source [QC] observes performance gain of 4.1% at CSI feedback overhead A (small overhead)
  + 1 source [QC] observes a performance gain of 0.4% at CSI feedback overhead B (medium overhead)
  + 1 source [QC] observes performance gain if 0.5% at CSI feedback overhead C (large overhead)
* For RU 40-69%, 1 sources [QC] observed performance gain of 2-5.6%
  + 1 sources [QC] observe performance gain of 5.6% at CSI feedback overhead A (small overhead)
  + 1 sources [QC] observe performance gain of 2% at CSI feedback overhead B (medium overhead)
  + 1 sources [QC] observe performance gain of 3.2% at CSI feedback overhead C (large overhead)
* For RU > 70%, 2 sources [QC] observes performance gain of 5.9% to 7.6%
  + 1 sources [QC] observed performance gain of 7% at CSI feedback overhead A (small overhead)
  + 1 sources [QC] observes performance gain of 5.9% at CSI feedback overhead B (medium overhead)
  + 1 sources [QC] observes performance gain of 7.6% at CSI feedback overhead C (large overhead)

For the evaluation of temporal domain aspects **Case 3** of AI/ML based CSI compression compared to the *benchmark in terms of 5% UPT under FTP* traffic, till RAN1 #118, for the mixed scenario of 80% indoor and 20% outdoor users:

For Max Rank 1,

* For RU <= 39%, 1 sources [CATT] observes performance gain of -0.6% to 9.3%%:
  + TBD performance gain at CSI feedback overhead A (small overhead)
  + 1 source [CATT] observes a performance gain of -0.6% at CSI feedback overhead B (medium overhead)
  + 1 source [CATT] observes performance gain if 9.3% at CSI feedback overhead C (large overhead)
* For RU 40-69%, 1 sources [QC] observed performance gain of 3.9-11.2%
  + 1 sources [QC] observe performance gain of 11.2% at CSI feedback overhead A (small overhead)
  + 1 sources [QC] observe performance gain of 3.9% at CSI feedback overhead B (medium overhead)
  + 1 sources [QC] observe performance gain of 7.5% at CSI feedback overhead C (large overhead)
* For RU > 70%, 1 sources [QC] observes performance gain of 19.8-27.3%
  + 1 sources [QC] observed performance gain of 19.8% at CSI feedback overhead A (small overhead)
  + 1 sources [QC] observes performance gain of 20.7% at CSI feedback overhead B (medium overhead)
  + 1 sources [QC] observes performance gain of 27.3% at CSI feedback overhead C (large overhead)

For Max Rank 2,

* For RU <= 39%, 1 sources [NTT Docomo] observes performance gain of 12.5-20.6%:
  + 1 source [NTT Docomo] observes performance gain of 20.6% at CSI feedback overhead A (small overhead)
  + 1 source [NTT Docomo] observes a performance gain of 12.5% at CSI feedback overhead B (medium overhead)
  + 1 source [NTT Docomo] observes performance gain if 12.9% at CSI feedback overhead C (large overhead)
* For RU 40-69%, 2 sources [QC, NTT Docomo] observed performance gain of 13.4-25.3%
  + 2 sources [QC, NTT Docomo] observe performance gain of 21-25.3% at CSI feedback overhead A (small overhead)
  + 2 sources [QC, NTT Docomo] observe performance gain of 14.6-21% at CSI feedback overhead B (medium overhead)
  + 2 sources [QC, NTT Docomo] observe performance gain of 13.4-18.9% at CSI feedback overhead C (large overhead)
* For RU > 70%, 2 sources [QC, NTT Docomo] observes performance gain of 27% to 51%
  + 2 sources [QC, NTT Docomo] observed performance gain of 37.2-50.7% at CSI feedback overhead A (small overhead)
  + 2 sources [QC, NTT Docomo] observes performance gain of 29.6-34.6% at CSI feedback overhead B (medium overhead)
  + 2 sources [QC, NTT Docomo] observes performance gain of 27.7-31% at CSI feedback overhead C (large overhead)

For Max Rank 4,

* For RU <= 39%, 1 sources [QC] observes performance gain of 13-16%:
  + 1 source [QC] observes performance gain of 15.6% at CSI feedback overhead A (small overhead)
  + 1 source [QC] observes a performance gain of 13.3% at CSI feedback overhead B (medium overhead)
  + 1 source [QC] observes performance gain if 15.7% at CSI feedback overhead C (large overhead)
* For RU 40-69%, 1 sources [QC] observed performance gain of 18-29%
  + 1 sources [QC] observe performance gain of 28.2% at CSI feedback overhead A (small overhead)
  + 1 sources [QC] observe performance gain of 18.5% at CSI feedback overhead B (medium overhead)
  + 1 sources [QC] observe performance gain of 21.6% at CSI feedback overhead C (large overhead)
* For RU > 70%, 2 sources [QC] observes performance gain of 25-33%
  + 1 sources [QC] observed performance gain of 32.2% at CSI feedback overhead A (small overhead)
  + 1 sources [QC] observes performance gain of 27.6% at CSI feedback overhead B (medium overhead)
  + 1 sources [QC] observes performance gain of 25.4% at CSI feedback overhead C (large overhead)

For the evaluation of temporal domain aspects **Case 3** of AI/ML based CSI compression compared to the *benchmark in terms of mean UPT* *under FTP* traffic, till RAN1 #118, for the scenario of 100% outdoor users:

For Max Rank 1,

* For RU <= 39%, 1 sources [Ericsson] observes performance gain of 0-1%:
  + 1 source [Ericsson] observes performance gain of 0% at CSI feedback overhead A (small overhead)
  + 1 source [Ericsson] observes performance gain of 1% at CSI feedback overhead B (medium overhead)
  + 1 source [Ericsson] observes performance gain of 1% at CSI feedback overhead C (large overhead)
* For RU of 40-69%, 1 sources [Ericsson] observes performance gain of 4-8%:
  + 1 source [Ericsson] observes performance gain of 4% at CSI feedback overhead A (small overhead)
  + 1 source [Ericsson] observes performance gain of 6% at CSI feedback overhead B (medium overhead)
  + 1 source [Ericsson] observes performance gain of 8% at CSI feedback overhead C (large overhead)
* For RU >= 70%, TBD performance gains

For the evaluation of temporal domain aspects **Case 3** of AI/ML based CSI compression compared to the *benchmark in terms of 5% UPT* *under FTP* traffic, till RAN1 #118, for the scenario of 100% outdoor users:

For Max Rank 1,

* For RU <= 39%, 1 sources [Ericsson] observes performance gain of -1% to 5%:
  + 1 source [Ericsson] observes performance gain of -1% at CSI feedback overhead A (small overhead)
  + 1 source [Ericsson] observes performance gain of 2% at CSI feedback overhead B (medium overhead)
  + 1 source [Ericsson] observes performance gain of 5% at CSI feedback overhead C (large overhead)
* For RU of 40-69%, 1 sources [Ericsson] observes performance gain of 8-17%:
  + 1 source [Ericsson] observes performance gain of 8% at CSI feedback overhead A (small overhead)
  + 1 source [Ericsson] observes performance gain of 7% at CSI feedback overhead B (medium overhead)
  + 1 source [Ericsson] observes performance gain of 17% at CSI feedback overhead C (large overhead)
* For RU >= 70%, TBD performance gains

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | Add the description to be aligned with other observations?  The above results are based on the following assumptions besides the assumptions of the agreed EVM table:  - Precoding matrix is used as the compression model input.  - Training data samples are not quantized, i.e., Float32 is used/represented.  - 1-on-1 joint training is assumed.  - Benchmark is Rel-18 doppler eT2. |
|  |  |

Observation 133a: Full buffer performance Case 3

For the evaluation of temporal domain aspects **Case 3** of AI/ML based CSI compression compared to the *benchmark, in terms of mean UPT under full buffer*, till RAN1 #118, for the mixed scenario of 80% indoor and 20% outdoor users,

* For Max rank 1,
  + 3 sources [QC, Xiaomi, Vivo] observes the performance gain of 1-11% at CSI feedback overhead A (small overhead);
  + 2 sources [QC, Xiaomi] observe performance gain of 2-5% at CSI feedback overhead B (medium overhead);
  + 2 sources [QC, Xiaomi] observe performance gain of 2-10% at CSI feedback overhead C (large overhead);
* For Max rank 2,
  + 4 sources [QC, Vivo, Fujitsu, ZTE] observes the performance gain of 1-19% at CSI feedback overhead A (small overhead);
  + 2 sources [QC, ZTE] observe performance gain of 5-7% at CSI feedback overhead B (medium overhead);
  + 2 sources [QC, ZTE] observe performance gain of 1-9% at CSI feedback overhead C (large overhead);
* For Max rank 4,
  + 1 source [QC] observes the performance gain of 7.3% at CSI feedback overhead A (small overhead);
  + 1 source [QC] observes performance gain of 6.3% at CSI feedback overhead B (medium overhead);
  + 1 source [QC] observes performance gain of 8.6% at CSI feedback overhead C (large overhead);

For the evaluation of temporal domain aspects **Case 3** of AI/ML based CSI compression compared to the *benchmark in terms of 5% UPT under full buffer*, till RAN1 #118, for the mixed scenario of 80% indoor and 20% outdoor users,

* For Max rank 1,
  + 3 sources [QC, Xiaomi, Vivo] observes the performance gain of 1-12% at CSI feedback overhead A (small overhead);
  + 2 sources [QC, Xiaomi] observe performance gain of 2-5.3% at CSI feedback overhead B (medium overhead);
  + 2 sources [QC, Xiaomi] observe performance gain of 4-13% at CSI feedback overhead C (large overhead);
* For Max rank 2,
  + 4 sources [QC, Vivo, Fujitsu, ZTE] observes the performance gain of 1-34% at CSI feedback overhead A (small overhead);
  + 2 sources [QC, ZTE] observe performance gain of 2-9% at CSI feedback overhead B (medium overhead);
  + 2 sources [QC, ZTE] observe performance gain of 1.5-9% at CSI feedback overhead C (large overhead);
* For Max rank 4,
  + 1 source [QC] observes the performance gain of 3% at CSI feedback overhead A (small overhead);
  + 1 source [QC] observes performance gain of 1.2% at CSI feedback overhead B (medium overhead);
  + 1 source [QC] observes performance gain of 7.5% at CSI feedback overhead C (large overhead);

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

- Precoding matrix of the current CSI is used as the model input.

- Training data samples are not quantized, i.e., Float32 is used/represented.

- 1-on-1 joint training is assumed.

- Benchmark is Rel-18 Type II codebook.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | Correct the description part as follows.  The above results are based on the following assumptions besides the assumptions of the agreed EVM table:  - Precoding matrix ~~of the current CSI~~ is used as the compression model input.  - Training data samples are not quantized, i.e., Float32 is used/represented.  - 1-on-1 joint training is assumed.  - Benchmark is Rel-18 Type II codebook. |
|  |  |

Observation 134a: CSI feedback reduction Case 3

For the evaluation of temporal domain aspects **Case 3** of AI/ML based CSI compression, compared to the benchmark, in terms of CSI feedback reduction, till RAN1 #118, for the mixed scenario of 80% indoor and 20% outdoor users,

* For Max rank 1,
  + For CSI feedback overhead A (small overhead),
    - 1 source [QC] observes CSI feedback reduction of 47% for full buffer.
    - CSI feedback reduction is TBD for FTP traffic at RU <= 39%
    - 1 source [QC] observes CSI feedback reduction of 0% for FTP traffic at RU of 40-69%
    - 1 source [QC] observes CSI feedback reduction of 29% for FTP traffic at RU >= 70%
  + For CSI feedback overhead B (medium overhead),
    - 1 source [QC] observes CSI feedback reduction of 29% for full buffer.
    - CSI feedback reduction is TBD for FTP traffic at RU <= 39%
    - 1 source [QC] observes CSI feedback reduction of 20% for FTP traffic at RU of 40-69%
    - 1 source [QC] observes CSI feedback reduction of 8% for FTP traffic at RU >= 70%
  + For CSI feedback overhead C (large overhead),
    - 1 source [QC] observes CSI feedback reduction of 57%
    - CSI feedback reduction is TBD for FTP traffic at RU <= 39%
    - 1 source [QC] observes CSI feedback reduction of 53% for FTP traffic at RU of 40-69%
    - 1 source [QC] observes CSI feedback reduction of 47% for FTP traffic at RU >= 70%
* For Max rank 2,
  + For CSI feedback overhead A (small overhead),
    - 1 source [QC] observes CSI feedback reduction of 39% for full buffer.
    - CSI feedback reduction is TBD for FTP traffic at RU <= 39%
    - 1 source [QC] observes CSI feedback reduction of 23% for FTP traffic at RU of 40-69%
    - 1 source [QC] observes CSI feedback reduction of 0% for FTP traffic at RU >= 70%
  + For CSI feedback overhead B (medium overhead),
    - 1 source [QC] observes CSI feedback reduction of 37% for full buffer.
    - CSI feedback reduction is TBD for FTP traffic at RU <= 39%
    - 1 source [QC] observes CSI feedback reduction of 11% for FTP traffic at RU of 40-69%
    - 1 source [QC] observes CSI feedback reduction of 0% for FTP traffic at RU >= 70%
  + For CSI feedback overhead C (large overhead),
    - 1 source [QC] observes CSI feedback reduction of 49%
    - CSI feedback reduction is TBD for FTP traffic at RU <= 39%
    - 1 source [QC] observes CSI feedback reduction of 20% for FTP traffic at RU of 40-69%
    - 1 source [QC] observes CSI feedback reduction of 30% for FTP traffic at RU >= 70%
* For Max rank 4,
  + For CSI feedback overhead A (small overhead),
    - 1 source [QC] observes CSI feedback reduction of 42% for full buffer.
    - 1 source [QC] observes CSI feedback reduction of 49% for FTP traffic at RU <= 39%
    - 1 source [QC] observes CSI feedback reduction of 52% for FTP traffic at RU of 40-69%
    - 1 source [QC] observes CSI feedback reduction of 43% for FTP traffic at RU >= 70%
  + For CSI feedback overhead B (medium overhead),
    - 1 source [QC] observes CSI feedback reduction of 40% for full buffer.
    - 1 source [QC] observes CSI feedback reduction of 5% for FTP traffic at RU <= 39%
    - 1 source [QC] observes CSI feedback reduction of 22% for FTP traffic at RU of 40-69%
    - 1 source [QC] observes CSI feedback reduction of 34% for FTP traffic at RU >= 70%
  + For CSI feedback overhead C (large overhead),
    - 1 source [QC] observes CSI feedback reduction of 51%
    - 1 source [QC] observes CSI feedback reduction of 12% for FTP traffic at RU <= 39%
    - 1 source [QC] observes CSI feedback reduction of 25% for FTP traffic at RU of 40-69%
    - 1 source [QC] observes CSI feedback reduction of 39% for FTP traffic at RU >= 70%

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | Add the description to be aligned with other observations?  The above results are based on the following assumptions besides the assumptions of the agreed EVM table:  - Precoding matrix is used as the compression model input.  - Training data samples are not quantized, i.e., Float32 is used/represented.  - 1-on-1 joint training is assumed.  - Benchmark is Rel-18 doppler eT2. |
|  |  |

### Evaluation results Case 4

Observation 141a: SGCS performance Case 4

For the evaluation of temporal domain aspects **Case 4** of AI/ML based CSI compression compared to the *benchmark in terms of SGCS*, till RAN1 #117, for the mixed scenario of 80% indoor and 20% outdoor users:

For Layer 1,

- 1 source [QC] observes the performance gain of 7.2% at CSI payload X (small payload);

- 2 sources [QC, ETRI] observe the performance gain of 3.61-4.7% at CSI payload Y (medium payload) ;

- 1 source [QC] observes the performance gain of 7.9% at CSI payload Z (large payload) .

For Layer 2,

- 1 source [QC] observes the performance gain of 22.9% at CSI payload X (small payload);

- 1 source [QC] observes the performance gain of 10.1% at CSI payload Y (medium payload) ;

- 1 source [QC] observes the performance gain of 16.6% at CSI payload Z (large payload) .

For Layer 3,

- 1 source [QC] observes the performance gain of 80.1% at CSI payload X (small payload);

- 1 source [QC] observes the performance gain of 26.2% at CSI payload Y (medium payload) ;

- 1 source [QC] observes the performance gain of 34% at CSI payload Z (large payload) .

For Layer 4,

- 1 source [QC] observes the performance gain of 104.1% at CSI payload X (small payload);

- 1 source [QC] observes the performance gain of 32.4% at CSI payload Y (medium payload) ;

- 1 source [QC] observes the performance gain of 38.2% at CSI payload Z (large payload) .

For the evaluation of temporal domain aspects **Case 4** of AI/ML based CSI compression compared to the *benchmark in terms of SGCS*, for the scenario of 100% outdoor users:

For Layer 1,

- 1 source [Apple] observes the performance gain of 10% at CSI payload X (small payload);

- ~~The performance gain at CSI payload Y (medium payload) is TBD;~~

~~- The performance gain at CSI payload Z (large payload) is TBD.~~

|  |  |
| --- | --- |
| *Company* | *Comments* |
|  |  |
|  |  |

### Evaluation results on inter-vendor training

Observation 161a: SGCS performance multi-vendor training

For the multi-vendor evaluation of temporal domain aspects **Case 1/2/3/4/5** of AI/ML based CSI compression assuming separate training, by comparing the performance with joint training baseline:

For Case 2

- 1 source [Huawei] observes the performance gain of 0.1% at CSI payload X (small payload);

- 2 sources [QC, Huawei] observe the performance gain of -0.86% to -0.12% at CSI payload Y (medium payload) ;

- 2 sources [QC, Huawei] observe the performance gain of -0.74% to -0.02% at CSI payload Z (large payload) .

For Case 3

- 1 source [QC] observes the performance gain of -0.65% to -0.53% at CSI payload Y (medium payload) ;

- 1 source [QC] observes the performance gain of -0.40% to -0.33% at CSI payload Z (large payload).

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Ericsson | More detailed assumptions made for separate training shall be stated in the observation. For instance,  • Is it one to N multi-vendor training, or N to M multi-vendor training, where N>1 and M >1?  • Is it NW first training or UE first training?  • Which inter-vendor training collaboration option is assumed?  • What information is assumed to be exchanged from one side to assist the training at the other side? |
|  |  |

### Others

Please provide any other comments for this section.

|  |  |
| --- | --- |
| *Company* | *Comments* |
|  |  |
|  |  |

# Localized models

## Summary of company proposals

From the submitted contributions, proposals related to the study of localized models, i.e., models specific to a cell, site, location, or region, are summarized below.

Tejas Networks

**Proposal 5: *In case of N different local regions and train N different localized models for each region, average performance should be considered over the N local regions*.**

ZTE

**Proposal 7*:*** *For cell/site specific model, prioritize the alignment on the understandings and EVMs for the cell/site specific model among companies first during Rel-19 study phase.*

**Proposal 8*:*** *For EVM calibration on cell/site specific model, prioritize to construct the dataset for cell/site specific model with the same number of data samples per cell/site compared with generalized model.*

OPPO

Proposal 7*: Suggest to study AI/ML based CSI compression with localized model in Rel-19, and discuss the EVM including the following aspects:*

* ***Impact of spatial consistency***
* ***Different scenarios, e.g., indoor/outdoor UE distributions, LoS/NLoS ratios.***

NVIDIA

**Proposal 1: *Site-specific AI/ML models for CSI compression should be considered to improve performance gain***

**Proposal 2: *Define a common reference scenario with site specificity as a basis for further study of AI/ML based CSI compression*.**

**Proposal 3: *Select one the following options to define a common reference scenario with site specificity as a basis for further study of AI/ML based CSI compression:***

* ***Option 1: Real-scenario map that is a virtual representation of a real area on earth.***
* ***Option 2: Synthetic-scenario map that is artificially constructed to mimic a certain environment such as urban macro, rural macro, indoor office, or indoor factory***

**Proposal 4: *With a common reference scenario with site specificity, ray tracing is used to generate channel data for the development and evaluation of site-specific AI/ML models for CSI compression.***

## Discussion

Observation 201a: SGCS performance and complexity of localized models Option 1

For the evaluation of AI/ML based CSI compression using localized models (Option 1), compared to the non-AI/ML *benchmark in terms of SGCS*, till RAN1 #118

For Layer 1,

* + 2 sources [ZTE, ViVo] observe the performance gain of 15-28% over benchmark, at CSI payload X (small payload), using TSF compression
  + 2 sources [ZTE, ViVo] observes the performance gain of 13-15 % at CSI payload Y (medium payload), using TSF compression
  + 2 sources [ZTE, ViVo] observes the performance gain of 8-11% at CSI payload Z (large payload), using TSF compression

For Layer 2,

* + 2 sources [ZTE, Vivo] observes a performance gain of 15-24% over benchmark, at CSI payload of X (small payload)
  + 2 sources [ZTE, Vivo] observes a performance gain of 16-21% over benchmark, at CSI payload of Y (medium payload)
  + 2 sources [ZTE, Vivo] observes a performance gain of 9-18% over benchmark, at CSI payload of Z (large payload)

For the evaluation of AI/ML based CSI compression using localized models (Option 1), compared to the AI/ML based CSI compression using global models *~~benchmark~~ in terms of SGCS*, where the localized models have the same complexity as the global models, till RAN1 #118,

For Layer 1,

* + 3 sources [ZTE, Vivo, QC] observes the performance gain of 1-11% over global model, at CSI payload X (small payload), using TSF compression
  + 3 sources [ZTE, Vivo, QC] observes the performance gain of 4-17% at CSI payload Y (medium payload), using TSF compression
  + 2 sources [ZTE, Vivo] observes the performance gain of 1-7% at CSI payload Z (large payload), using TSF compression

For Layer 2,

* + 2 sources [ZTE, Vivo] observes a performance gain of 5-14% over global model, at CSI payload of X (small payload)
  + 2 sources [ZTE, Vivo] observes a performance gain of 5-19% over global model, at CSI payload of Y (medium payload)
  + 2 sources [ZTE, Vivo] observes a performance gain of 0-13% over global model, at CSI payload of Z (large payload)

For the evaluation of AI/ML based CSI compression using localized models (Option 1), compared to the AI/ML based CSI compression using global models *~~benchmark~~ in terms of SGCS*, where the localized models have the reduced complexity as the global models, till RAN1 #118,

For Layer 1,

* + 1 source [QC] observes the performance gain of 1.6% over global model, with 27% of parameters and 1% of FLOPs as the global model, at CSI payload X (small payload)
  + 1 source [QC] observes the performance gain of 11.9% over the global model, with 27% of parameters and 1% of FLOPs as the global model,  at CSI payload Y (medium payload)
  + ~~TBD performance gain at CSI payload Z (large payload)~~

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

* Precoding matrix of the current CSI is used as the model input.
* Training data samples are not quantized, i.e., Float32 is used/represented.
* 1-on-1 joint training is assumed.
* The performance metric is SGCS for Layer 1 of Max rank 1 or Layer 1/2 of Max rank 2.
* CSI payload X is ≤ 80 bits; CSI payload Y is 100 bits - 140 bits; CSI payload Z is ≥ 230 bits; where X, Y, Z are applicable per layer.
* Benchmark is Rel-16 Type II codebook.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | From our review of Table X3, we understand that some source does not adopt TSF compression, which seems to appear in each sub-bullet. |
| Intel | Option 1 and 2 are both applicable for Intel’s results: global model was trained for 100% outdoor NLOS UEs without spatial consistency, while local model was trained for 100% outdoor NLOS UEs per each cell (sector) of one site with spatial consistency.  Considering the above, is it possible to capture our results in the observation for Option 1 as well? |
| ZTE | I have a question about the proposal since we didn’t provide the SGCS performance of TSF compression, just for SF compression like Rel-18 compression. So, maybe a clarification is needed how SF compression is captured to differentiate the Rel-19 TSF cases and Rel-18 SF case. |
|  |  |
|  |  |

Observation 202a: SGCS performance and complexity of localized models Option 2

For the evaluation of AI/ML based CSI compression using localized models (Option 2), compared to the non-AI/ML *benchmark in terms of SGCS*, till RAN1 #118

For Layer 1,

* + 3 sources [ViVo, Oppo, Intel] observe a performance gain of 4.5-10% over benchmark, at CSI payload of X (small payload)
  + Performance gain of TBD % over benchmark, at CSI payload of Y (medium payload)
  + Performance gain of TBD % over benchmark, at CSI payload of Y (medium payload)

For the evaluation of AI/ML based CSI compression using localized models (Option 2), compared to the AI/ML based CSI compression using global models *~~benchmark~~ in terms of SGCS*, where the localized models have the same complexity as the global models, till RAN1 #118,

For Layer 1,

* + 6 sources [Nokia, Vivo, Panasonic, Oppo, Futurewei, Intel] observes a performance gain of -2.65% to 6% over global model, at CSI payload of X (small payload)
  + 1 source [Nokia] observes a performance gain of 0-2% over global model, at CSI payload of Y (medium payload)
  + 1 source [Nokia] observes a performance gain of -0.5% to 3% over global model, at CSI payload of Y (medium payload)

For the evaluation of AI/ML based CSI compression using localized models (Option 2), compared to the AI/ML based CSI compression using global models *~~benchmark~~ in terms of SGCS*, where the localized models have the lower complexity as the global models, till RAN1 #118,

For Layer 1,

* + 1 sources [Intel] observes a performance gain of 0% over global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI payload of X (small payload)
  + TBD performance gain over global model, at CSI payload of Y (medium payload)
  + TBD performance gain over global model, at CSI payload of Y (medium payload)

The above results are based on the following assumptions besides the assumptions of the agreed EVM table:

* Precoding matrix of the current CSI is used as the model input.
* Training data samples are not quantized, i.e., Float32 is used/represented.
* 1-on-1 joint training is assumed.
* The performance metric is SGCS for Layer 1 of Max rank 1 or Layer 1/2 of Max rank 2.
* CSI payload X is ≤ 80 bits; CSI payload Y is 100 bits - 140 bits; CSI payload Z is ≥ 230 bits; where X, Y, Z are applicable per layer.
* Benchmark is Rel-16 Type II codebook.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Panasonic | We are fine with the proposed observation. Our results are correctly captured. |
|  |  |

Observation 203a: FTP performance of localized models Option 2

For the evaluation of localized models of AI/ML based CSI compression compared to the *benchmark, in terms of mean UPT under FTP traffic*, till RAN1 #118,

* For Max Rank 1,
* For RU <= 39%, 1 source [Intel] observes 0% performance gain
* For RU of 40-69%, 1 source [Intel] observes 0.1% performance gains
* For RU >= 70%, 1 source [Intel] observes 0.9% performance gains

For the evaluation of localized models of AI/ML based CSI compression compared to compression using global models, *in terms of mean UPT under FTP traffic,* where localized models have the same complexity as the *global model,* till RAN1 #118,

* For Max Rank 1,
* For RU <= 39%, 1 source [Intel] observes 0% performance gain
* For RU of 40-69%, 1 source [Intel] observes 0.2% performance gains
* For RU >= 70%, 1 source [Intel] observes 1.2% performance gains

For the evaluation of localized models of AI/ML based CSI compression compared to compression using global models, *in terms of mean UPT under FTP traffic,* where localized models have lower complexity than the *global model,* till RAN1 #118,

* For Max Rank 1,
* For RU <= 39%, 1 source [Intel] observes 0% performance gain over the global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI feedback overhead A (small overhead).
* For RU of 40-69%, 1 source [Intel] observes 0.3% performance gains over the global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI feedback overhead A (small overhead)
* For RU >= 70%, 1 source [Intel] observes 0.9% performance gains over the global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI feedback overhead A (small overhead)

For the evaluation of localized models of AI/ML based CSI compression compared to the *benchmark, in terms of 5% UPT under FTP traffic*, till RAN1 #118,

* For Max Rank 1,
* For RU <= 39%, 1 source [Intel] observes -0.4% performance gain
* For RU of 40-69%, 1 source [Intel] observes 0.6% performance gains
* For RU >= 70%, 1 source [Intel] observes 0.6% performance gains

For the evaluation of localized models of AI/ML based CSI compression compared to compression using global models, *in terms of 5% UPT under FTP traffic,* where localized models have the same complexity as the *global model,* till RAN1 #118,

* For Max Rank 1,
* For RU <= 39%, 1 source [Intel] observes -0.6% performance gain
* For RU of 40-69%, 1 source [Intel] observes 0.8% performance gains
* For RU >= 70%, 1 source [Intel] observes 2.5% performance gains

For the evaluation of localized models of AI/ML based CSI compression compared to compression using global models, *in terms of 5% UPT under FTP traffic,* where localized models have lower complexity than the *global model,* till RAN1 #118,

* For Max Rank 1,
* For RU <= 39%, 1 source [Intel] observes 0.4% performance gain over the global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI feedback overhead A (small overhead)
* For RU of 40-69%, 1 source [Intel] observes 2.0% performance gains over the global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI feedback overhead A (small overhead)
* For RU >= 70%, 1 source [Intel] observes 3.8% performance gains over the global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI feedback overhead A (small overhead).

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | Do we need to draw the observation if only a single source provides results? |
| Intel | Change is needed for gains of low complexity local model comparing to global model according to Section 2.3 of this document.  *For the evaluation of localized models of AI/ML based CSI compression compared to compression using global models, in terms of mean UPT under FTP traffic, where localized models have lower complexity than the global model, till RAN1 #118,*   * *For Max Rank 1,* * *For RU <= 39%, 1 source [Intel] observes -0.1% performance gain over the global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI feedback overhead A (small overhead).* * *For RU of 40-69%, 1 source [Intel] observes -0.5% performance gains over the global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI feedback overhead A (small overhead)* * *For RU >= 70%, 1 source [Intel] observes -1.4% performance gains over the global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI feedback overhead A (small overhead)*   *For the evaluation of localized models of AI/ML based CSI compression compared to compression using global models, in terms of 5% UPT under FTP traffic, where localized models have lower complexity than the global model, till RAN1 #118,*   * *For Max Rank 1,* * *For RU <= 39%, 1 source [Intel] observes 0.2% performance gain over the global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI feedback overhead A (small overhead)* * *For RU of 40-69%, 1 source [Intel] observes -2.4% performance gains over the global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI feedback overhead A (small overhead)* * *For RU >= 70%, 1 source [Intel] observes -4.1% performance gains over the global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI feedback overhead A (small overhead).* |

Observation 205a: Full buffer performance of localized models Option 1

For the evaluation of AI/ML based CSI compression using localized models (Option 1), compared to the *benchmark, in terms of mean UPT under full buffer*, till RAN1 #118,

* For Max Rank 1, TBD performance gains
* For Max Rank 2,
* 1 source [Vivo] observes performance gain of 10.8-15.2%,
  + 1 source [Vivo] observes the performance gain of 15.2% at CSI feedback overhead A (small overhead);
  + 1 source [Vivo] observes the performance gain of 10.8% at CSI feedback overhead B (medium overhead);
  + 1 source [Vivo] observes the performance gain of 11.9% at CSI feedback overhead A (small overhead)
* For Max Rank 4, TBD performance gains.

For the evaluation of AI/ML based CSI compression using localized models (Option 2), compared to the compression using global models *~~benchmark~~ in terms of mean UPT under full buffer*, where the localized models have the same complexity as the global models, till RAN1 #118,

* For Max Rank 1, TBD performance gains
* For Max Rank 2,
* 1 source [Vivo] observes performance gain of 3.87-6.43%,
  + 1 source [Vivo] observes the performance gain of 3.87% at CSI feedback overhead A (small overhead);
  + 1 source [Vivo] observes the performance gain of 6.43% at CSI feedback overhead B (medium overhead);
  + 1 source [Vivo] observes the performance gain of 6.36% at CSI feedback overhead A (small overhead)
* For Max Rank 4, TBD performance gains.

For the evaluation of AI/ML based CSI compression using localized models (Option 1), compared to the *benchmark, in terms of 5% UPT under full buffer*, till RAN1 #118,

* For Max Rank 1, TBD performance gains
* For Max Rank 2,
* 1 source [Vivo] observes performance gain of 12.8-20.5%,
  + 1 source [Vivo] observes the performance gain of 17.6% at CSI feedback overhead A (small overhead);
  + 1 source [Vivo] observes the performance gain of 12.8% at CSI feedback overhead B (medium overhead);
  + 1 source [Vivo] observes the performance gain of 20.5% at CSI feedback overhead A (small overhead)
* For Max Rank 4, TBD performance gains.

For the evaluation of AI/ML based CSI compression using localized models (Option 2), compared to the compression using global models *~~benchmark~~ in terms of 5% UPT under full buffer*, where the localized models have the same complexity as the global models, till RAN1 #118,

* For Max Rank 1, TBD performance gains
* For Max Rank 2,
* 1 source [Vivo] observes performance gain of 2.52-9.83%,
  + 1 source [Vivo] observes the performance gain of 2.52% at CSI feedback overhead A (small overhead);
  + 1 source [Vivo] observes the performance gain of 9.83% at CSI feedback overhead B (medium overhead);
  + 1 source [Vivo] observes the performance gain of 9.24% at CSI feedback overhead A (small overhead)
* For Max Rank 4, TBD performance gains.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | Do we need to draw the observation if only a single source provides results? |
|  |  |

Observation 206a: Full buffer performance of localized models Option 2

For the evaluation of AI/ML based CSI compression using localized models (Option 2), compared to the *benchmark, in terms of mean UPT under full buffer*, till RAN1 #118,

* For Max Rank 1,
* 1 source [Intel] observes performance gain of 3.7%,
  + 1 source [Intel] observes the performance gain of 3.7% at CSI feedback overhead A (small overhead);
  + TBD performance gain at CSI feedback overhead B (medium overhead)
  + TBD performance gain at CSI feedback overhead C (large overhead)
* For Max Rank 2, TBD performance gains
* For Max Rank 4, TBD performance gains.

For the evaluation of AI/ML based CSI compression using localized models (Option 2), compared to the compression using global models *~~benchmark~~ in terms of mean UPT under full buffer*, where the localized models have the same complexity as the global models, till RAN1 #118

* For Max Rank 1,
* 1 source [Intel] observes performance gain of 3.8%,
  + 1 source [Intel] observes the performance gain of 3.8% at CSI feedback overhead A (small overhead);
  + TBD performance gain at CSI feedback overhead B (medium overhead)
  + TBD performance gain at CSI feedback overhead C (large overhead)
* For Max Rank 2, TBD performance gains
* For Max Rank 4, TBD performance gains.

For the evaluation of AI/ML based CSI compression using localized models (Option 2), compared to the compression using global models *~~benchmark~~ in terms of mean UPT under full buffer*, where the localized models have lower complexity as the global models, till RAN1 #118,

* For Max Rank 1,
* 1 source [Intel] observes performance gain of 2.7%,
  + 1 source [Intel] observes the performance gain of 2.7% over the global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI feedback overhead A (small overhead);
  + TBD performance gain at CSI feedback overhead B (medium overhead)
  + TBD performance gain at CSI feedback overhead C (large overhead)
* For Max Rank 2, TBD performance gains
* For Max Rank 4, TBD performance gains.

For the evaluation of localized models of AI/ML based CSI compression compared to the *benchmark, in terms of 5% edge UPT under full buffer*, till RAN1 #118,

* For Max Rank 1,
* 1 source [Intel] observes performance gain of -2.0%,
  + 1 source [Intel] observes the performance gain of -2% at CSI feedback overhead A (small overhead);
  + TBD performance gain at CSI feedback overhead B (medium overhead)
  + TBD performance gain at CSI feedback overhead C (large overhead)
* For Max Rank 2, TBD performance gains
* For Max Rank 4, TBD performance gains.

For the evaluation of localized models of AI/ML based CSI compression compared compression using global models*, in terms of 5% edge UPT under full buffer*, where the localized models have same complexity as the global models, till RAN1 #118,

* For Max Rank 1,
* 1 source [Intel] observes performance gain of 2.2%,
  + 1 source [Intel] observes the performance gain of 2.2% at CSI feedback overhead A (small overhead);
  + TBD performance gain at CSI feedback overhead B (medium overhead)
  + TBD performance gain at CSI feedback overhead C (large overhead)
* For Max Rank 2, TBD performance gains
* For Max Rank 4, TBD performance gains.

For the evaluation of localized models of AI/ML based CSI compression compared to compression using global models*, in terms of 5% edge UPT under full buffer*, where the localized models have lower complexity as the global models, till RAN1 #118,

* For Max Rank 1,
* 1 source [Intel] observes performance gain of 3.4%,
  + 1 source [Intel] observes the performance gain of 3.4% over the global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI feedback overhead A (small overhead);
  + TBD performance gain at CSI feedback overhead B (medium overhead)
  + TBD performance gain at CSI feedback overhead C (large overhead)
* For Max Rank 2, TBD performance gains
* For Max Rank 4, TBD performance gains.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | Do we need to draw the observation if only a single source provides results? |
| Intel | Change is needed for gains of low complexity local model comparing to global model according to Section 2.3 of this document.  *For the evaluation of AI/ML based CSI compression using localized models (Option 2), compared to the compression using global models ~~benchmark~~ in terms of mean UPT under full buffer, where the localized models have lower complexity as the global models, till RAN1 #118,*   * *For Max Rank 1,* * *1 source [Intel] observes performance gain of -1.2%,*   + *1 source [Intel] observes the performance gain of -1.2% over the global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI feedback overhead A (small overhead);*   + *TBD performance gain at CSI feedback overhead B (medium overhead)*   + *TBD performance gain at CSI feedback overhead C (large overhead)* * *For Max Rank 2, TBD performance gains* * *For Max Rank 4, TBD performance gains.*   *For the evaluation of localized models of AI/ML based CSI compression compared to compression using global models, in terms of 5% edge UPT under full buffer, where the localized models have lower complexity as the global models, till RAN1 #118,*   * *For Max Rank 1,* * *1 source [Intel] observes performance gain of -3.0%,*   + *1 source [Intel] observes the performance gain of -3.0% over the global model, with 2.8% of parameters and 2.6% of FLOPs as the global model, at CSI feedback overhead A (small overhead);*   + *TBD performance gain at CSI feedback overhead B (medium overhead)*   + *TBD performance gain at CSI feedback overhead C (large overhead)* * *For Max Rank 2, TBD performance gains*   *For Max Rank 4, TBD performance gains.* |

### Others

Please provide any other comments for this section.

|  |  |
| --- | --- |
| *Company* | *Comments* |
|  |  |
|  |  |

# Inter-vendor training collaboration

## Summary of company proposals

From the submitted contributions, proposals related to inter-vendor training collaboration are summarized below.

Futurewei

***Proposal 1: Among the (sub) options to alleviate/resolve the issues related to inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model, consider using standardized signalling in upper layers (case z4) as one of the delivery options to save the air-interface overhead at least for Option 3a/4/5a.***

***Proposal 2: For AI/ML-based CSI compression using two-sided model, further study the following potential specification impact related to quantization of CSI feedback, at least for Option 3a/3b/4/5a in alleviating/resolving the issues related to inter-vendor training collaboration:***

* ***Vector quantization:***
  + ***Exchange of vector quantization codebook(s).***
  + ***Segmentation information (if segmentation is used) of the CSI output.***
* ***Scalar quantization:***
  + ***Configuration of quantization granularity and the corresponding range values.***
  + ***Exchange of scalar quantization dictionary.***

Huawei

***Proposal 4: Capture Table 8 to the TR 38.843 for the comparison over the model exchange related options/sub-options of inter-vendor training collaboration.***

Table 8 Comparison over model exchange related options/sub-options

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model exchange related options/sub-options** | | | **Inter-vendor collaboration complexity** | **Performance** | **Feasibility/potential spec impacts** | **Exchange overhead** | **Offline training complexity** |
| Option 3  Parameters exchange | 3a  With offline engineering | 3a-1 | Relieved if the parameter exchange is performed with standardized procedure | More limited than 3b | Depends on feasibility of model transfer. Non-trivial spec effort/spec evolution | Medium | Medium |
| 3a-2 | Medium | High |
| 3a-3 | High | Medium or high |
| 3b  Directly inference | | Partially limited | Low | Low |
| Option 5  Model exchange | 5a  With offline engineering | 5a-1 | Not relieved.  Need offline model structure alignment | More limited than 5b | Medium | Medium |
| 5a-2 | Medium | High |
| 5a-3 | High | Medium or high |
| 5b  Directly inference | | Less limited | Low | Low |
| a: 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.  b: directly used for inference at the UE without offline engineering, potentially with on-device operations. | | | | | | | |

***Proposal 5: Capture Table 10 to the TR 38.843 for the comparison over the dataset exchange related sub-options of inter-vendor training collaboration.***

Table 10 Comparison over dataset exchange related method sub-options

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset exchange related sub-options** | | **Inter-vendor collaboration complexity** | **Performance** | **Feasibility/potential spec impacts** | **Exchange overhead** | **Offline training complexity** |
| Option 4  Parameters transfer, standardized data format | 4-1 | Relieved if dataset exchange is performed with standardized procedure | Optimal | Depends on whether dataset delivery is achieved with low collaboration complexity. Less spec effort than other options. | Low | Medium |
| 4-2 | Limited | Low | Medium or high depending on the training method |
| 4-3 | Optimal | High | Medium or high depending on the training method |

***Observation 10: RAN1 may need more time/effort to justify the feasibility for standardizing Option 1/3/4/5, considering:***

* ***For Option 1/3, the discussions leading to an exact model/model structure to be standardized have not started and are expected to be time consuming.***
* ***For Option 4/5, the discussions of specific dataset format and model representation format have not started.***
* ***For Option 3/4/5, the specific path (over-the-air or other approaches) and corresponding solutions for how to deliver parameter/dataset/model have not been sufficiently investigated.***
* ***Further down selection among options and sub-options may be needed to limit the scope of the study.***

Spreadtrum, BUPT

***Observation 4: For options 3 and option 5,*** ***how to define and standardize the model format needs further consideration.***

***Observation 5: For options 4,*** ***how to the content of data/dataset format should be further considered.***

Google

***Proposal 13: Consider to prioritize the option 3 and option 5 for inter-vendor training collaboration for further study.***

* ***Whether the UE directly uses the received parameter/model or not is up to UE implementation***

Tejas Networks

**Proposal 6: Consider the following as an inter-vendor complexity and feasibility for Option 3a-1/5a-1: Model exchange is required for both CSI generation and CSI reconstruction, so for CSI generation, parameters exchanged are preferred compared to the model exchange.**

**Proposal 7:** **Consider the following as an inter-vendor complexity and feasibility for option 3a-2/5a-2.**

* **Option 3a-2/5a-2: Similarly, as in Option 3a-1/5a-1, parameters exchanged are preferred for CSI reconstruction compared to the model exchange.**

**Proposal 8: Consider the following as an inter-vendor complexity and feasibility for option 3a-3/5a-3.**

* **Option 3a-3/5a-3: The model exchange is preferred for both CSI generation and CSI reconstruction, but a standardised reference model is needed for inter-vendor compatibility.**

**Proposal 9: Consider the following as an inter-vendor complexity for option 3b**

* **Option 3b: The method of exchanging of model/parameter over the air-interface via model transfer/delivery Case z4 will be more inter-vendor compatible with extra link overhead compared to 3a/5a.**

**Proposal 10: Consider the following as an inter-vendor complexity for option 5b**

* **Option 5b: The method of exchanging over the air interface via model transfer/delivery Case z4, it may not be inter-vendor compatible as the standardised model format may not follow any standard reference model.**

**Proposal 11: Consider the following as an inter-vendor complexity for option 4**

* **Option 4-1: Dataset (target CSI, CSI feedback) exchange from NW side and UE side (despite standardised data/dataset format) over the air interference requires huge link overhead. However, this is an essential task, so it needs to find an offline mechanism to exchange the data to UE and train the model which is placed at the UE. The same is applicable for options 4-2 and 4-3.**

CMCC

*Proposal 1: The progress on reference model in RAN4 could be reused for the standardization of reference model structure in Option 3, no matter the standardized model itself or the method on how to standardize an AI model.*

*Proposal 2: It should be assumed the CSI generation or reconstruction part that exchanged parameters belongs to is aligned with the CSI generation or reconstruction part that will be standardized in Option 3:*

* *For Option 3a-1,* *at least the reference model structure of CSI generation part should be standardized*
* *For Option 3a-2, at least* *the reference model structure of CSI reconstruction part should be standardized*
* *For Option 3a-3, the reference model structure of CSI generation part and reconstruction part should be standardized*
* *For Option 3b, at least the reference model structure of CSI generation part should be standardized*

*Proposal 3: There might be the following possible alternatives for model deployment procedure of Option 3a-1:*

* *Alt 1: The applied CSI generation part consists of reference model structure via pre-deployment and model parameters through offline engineering.*
* *Alt 2: The applied CSI generation part consists of model structure and parameters through offline engineering.*

*Proposal 4: There might be the following possible alternatives for model deployment procedure of Option 3a-2:*

* *Alt 1: The applied CSI generation part consists of reference model structure via pre-deployment and model parameters through offline engineering.*
  + *The reference model structure of both CSI generation part and reconstruction part should be standardized.*
* *Alt 2: The applied CSI generation part consists of model structure and parameters through offline engineering.*

*Proposal 5: There might be the following possible alternatives for model deployment procedure of Option 3a-3:*

* *Alt 1: The applied CSI generation part consists of reference model structure via pre-deployment and model parameters through offline engineering.*
  + *The standardized reference model structure of CSI generation part is applied for actual model deployment.*
* *Alt 2: The applied CSI generation part consists of model structure and parameters through offline engineering.*
  + *The standardized reference model structure of CSI generation part is only used for exchanged model parameters alignment.*

*Observation 9: Option 3a need UE have the following capabilities:*

* ***The capability to support the standardized model structure;***
* ***The capability to access the UE-side OTT server, including upload and download the model related information;***
* ***The capability to reload the updated model from the perspective of chipset;***
* ***The capability to perform parameters exchange either over the air-interface or offline;***

*Observation 10: Option 3b need UE have the following capabilities:*

* ***The capability to support the standardized model structure;***
* ***Good model scalability over different parameters during UE implementation, including chipset design and model design;***
* ***The capability to perform parameters exchange over the air-interface;***

*Proposal 6: It is proposed to discuss the inter-vendor collaboration complexity, performance and feasibility of Option 4-1, 4-2 and 4-3 together.*

*Observation 11: Option 4 need UE have the following capabilities:*

* ***The capability to access the UE-side OTT server, including upload received dataset and potential additional information, and download the trained model;***
* ***The capability to reload the updated model from the perspective of chipset;***
* ***The capability to perform parameters exchange either over the air-interface or offline;***

*Proposal 7: It should be assumed the CSI generation or reconstruction part exchanged is aligned with the CSI generation or reconstruction part of which the model structure aligned based on offline inter-vendor collaboration in Option 5:*

* *For Option 5a-1, at least the model structure of CSI generation part should be aligned offline*
* *For Option 5a-2, at least the model structure of CSI reconstruction part should be aligned offline*
* *For Option 5a-3, the model structure of CSI generation part and reconstruction part should be aligned offline*
* *For Option 5b, at least the model structure of CSI generation part should be aligned offline*

*Proposal 8: It is proposed to discuss the inter-vendor collaboration complexity, performance and feasibility of Option 5a-1, 5a-2, and 5a-3 together.*

*Observation 12: Option 5a need more time for model deployment compared to Option 5b.*

*Observation 13: The performance of Option 5a may be better than Option 3b.*

*Observation 14: The needed UE capabilities or UE chipset implementation over Option 5a and Option 5b are different.*

*Observation 15: The multi-vendor collaboration issue may be relieved for Option 5a if model exchange is performed over the air.*

*Observation 16: The multi-vendor collaboration issue still exists for Option 5b.*

*Observation 17: Both Option 5a and Option 5b refer to model transfer/delivery Case z4.*

*Observation 18: Option 5a need UE have the following capabilities:*

* ***The capability to access the UE-side OTT server, including upload and download the model related information;***
* ***The capability to reload the updated model from the perspective of chipset;***
* ***The capability to perform parameters exchange either over the air-interface or offline;***

*Observation 19: Option 5b need UE have the following capabilities:*

* ***Good model scalability over different parameters during UE implementation, including chipset design and model design;***
* ***The capability to perform parameters exchange over the air-interface;***

*Proposal 9: At least Option 3a, 3b and 5a should be prioritized for further study.*

Intel

***Proposal 2***:

* *For Options 3a and 5a, consider support of offline and over the air interface transfer/delivery following, e.g., model transfer/delivery Cases y and z4 respectively.*

***Observation 5***:

* *Compared to Options 3b/5b, Options 3a/5a can be expected to incur inherent additional latency and burden/efforts that can impact their practical applicability/responsiveness in context of LCM.*
* *At least for Option 3 family that involves specified model structure(s) with parameters being transferred from NW-side to UE-side, it may not be essential for received parameters from NW-side to go through offline engineering for re-training/re-development.*

***Proposal 3***:

* *At least for Option 3 for inter-vendor collaboration, Option 3b is prioritized over Options 3a.*

***Proposal 4***:

* *For Options 3a/5a, provision of dataset or information related to collecting data are not considered further in the context of inter-vendor collaboration for two-sided models.*
* *Note: This does not imply that provision of dataset or information related to collecting data are precluded.*

***Proposal 5***:

* *If performance targets are provided by NW-side to UE-side* *to help UE-side offline engineering and/or on-device adjustments and provide performance guidance, they are interpreted as assistance information and it is not expected that the UE should achieve the corresponding targets.*

***Proposal 6***:

* *Towards providing assistance to UE-side* *to help UE-side offline engineering and/or on-device adjustments and provide performance guidance, further study options for NW-side to provide targets/guidance on “CSI quality” represented using suitable metrics (e.g., NMSE, SGCS, etc.) and thresholds/targets on CSI accuracy performance.*
  + *Alternatively, or additionally, for certain specified or configured CSI quality metrics, a UE could be configured to report the quality of the compressed CSI that the NW may further utilize for model/functionality LCM.*

***Proposal 7***:

* *RAN1 to discuss mapping of different options for training collaboration agreed at RAN1#116 and training collaboration types assumed for UE/NW part training used in the actual operation at UE/NW side.*
  + *Consider the below table as a starting point for the discussion.*
  + *Note: Transfer/delivery from UE-side to NW-side are not listed below for compactness.*

ZTE

***Proposal 9:*** *For Option 1, prioritize to study the case of standardized CSI generation part.*

***Observation 23:*** *For Option 3a with offline secondary development, interoperability issue may be resolved without offline additional information exchange. However, the actual end-to-end performance after model deployment may not be guaranteed if either side and/or both sides perform offline secondary development before actual system use.*

***Observation 24:*** *Performance of Option 3a may be guaranteed via offline engineering with exchanged additional information. However, the problem of proprietary information disclosure would hinder the interoperability issue between UE side and NW side.*

***Observation 25:*** *Though the interoperability issue of Option 3a-3 may be resolved, the end-to-end performance is limited to the whole specified two-sided reference model structure, which may have lower performance upper-bound compared with Option 3a-1 and Option 3a-2.*

***Observation 26:*** *Compared with Option 3a, Option 3b experiences shorter timescale to do model inference without offline engineering if the format of the received parameters can be interpreted and compiled by UE device.*

***Proposal 10:*** *For Option 3, further study and evaluate the actual end-to-end performance of different sub-options when either side and/or both sides perform offline secondary development before actual system use.*

***Observation 27:*** *(option 4) The interoperability issue is questionable when model backbones are misaligned between UE side and NW side.*

***Observation 28:*** *(Option 4) Over-the-air dataset exchange would consume huge resource overhead.*

***Observation 29:*** *Option 5 shares the same conclusion of interoperability issue as Option 3, however, Option 5 incurs more offline additional inter-vendor collaboration complexity for reference model structure alignment.*

***Proposal 11:*** *For Option 3/4/5, prioritize to study NW-first training scheme and exchanging the parameters/model/dataset from NW side to UE side.*

***Proposal 12:*** *For Option 3/4/5, prioritize to study over-the-air delivery scheme.*

***Proposal 13:*** *For Option 3a/4/5a, further study the post-development operations on the AI/ML model during the actual deployment to guarantee the interoperability issue and performance when offline engineering is performed by UE vendors and/or NW vendors.*

***Proposal 14****: For comprehensive analysis on AI/ML framework for two-sided model, further study a complete and unified solution for model identification, multi-vendor collaboration, and model pairing.*

Ericsson

1. RAN1 should study and conclude that based on the standardized reference model in RAN1 Option 1 (RAN4 Option 3 or RAN4 Option 4), UE side and NW side can independently further improve their actual UE-part model and NW-part model using dataset collected in the field, if needed.
2. For inter-vendor training collaboration, UE-side first training cases for RAN1 Option 3/4/5, where the exchange of parameter/dataset/model originating from the UE-side and ending at the NW-side is not supported as it necessitates multiple models to operate in parallel at the NW-side.
3. Option 5a for inter-vendor training collaboration is not supported.

Observation 5: For Option 3a-2/3, it is not feasible for the NW-side to provide actual reconstructed model parameters to the UE-side, since it risks the NW-side to disclose its proprietary implementation related information.

1. Option 3a-2/3 for inter-vendor training collaboration is not supported.
2. For Option 3a, Over-the-air delivery method for exchanging information from the NW-side to UE-side is not supported.
3. For Option 3b, to mitigate inter-vendor collaboration complexity and improve feasibility for UE to use received parameters directly for inference, at least the model parameter precision and input data pre-processing must be standardized together with the CSI generation model structure. RAN1 should conclude on the required additional information that needs to be standardized to enable option 3b.
4. For Option 3b, the following aspects needs to be studied and concluded:
   * The feasibility and complexity of standardizing the CSI generation model structure.
   * The feasibility and complexity of standardizing the model parameter precision and input data pre-processing for the CSI generation model.
   * Any additional information that needs to be standardized to improve the feasibility for a UE to use received parameters directly for inference?
5. Option 5b for inter-vendor training collaboration is not supported.
6. For option 4, conclude that the UE-side training should be assisted by additional information about the NW side training/testing procedure:
   1. Any data preprocessing steps applied to the input of the encoder.
   2. Training objectives:
      1. Loss function that aligns with the decoder’s requirements
      2. If the NW side expects that the training and the corresponding latent space should capture certain features or distribution.
      3. Evaluation metric including how well the latent space aligns with the expected input to the decoder, and how the reconstruction accuracy should be.
   3. Postprocessing steps applied to the output of the encoder, e.g. if the encoder output is regularized or constrained to match this distribution.
   4. Potentially, the following could also help improve the performance:
      1. Model architecture, optimization strategies, regularization (e.g. drop out or added noise)
   5. Validation/testing data to assess compatibility.
   6. Performance targets include e.g. reconstruction accuracy, intermediate KPI like SGCS/NMSE/loss-value, and latent space validation.
7. For Option 4, over-the-air delivery method for exchanging information from the NW-side to UE-side is not supported.
8. Option 4-2 and 4-3 for inter-vendor training collaboration is not supported.

Vivo

1. **RAN1 considers specifying the reference model structure as a starting point for the specification of a reference model.**
2. **Towards the specification of reference model structure, the following procedures can be considered in RAN1:**
   * + - **Step 0: Aligning evaluation assumptions**
       - **Step 1: Determine the model backbone based on consensus and evaluation results on complexity and performance.**
       - **Step 2: Determine the model hyperparameters that need to be aligned.**
       - **Step 3: Align the hyperparameters of the model.**
3. **Characteristics of options to alleviate / resolve the issues related to inter-vendor training collaboration of AI/ML-based CSI compression can be summarized as:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Inter-vendor collaboration complexity** | **Performance** | **Interoperability and RAN4 / testing related aspects** | **Feasibility** | **Deployment timescale** |
| **Option 1** | Minimum complexity | Restricted | Solved | feasible | \ |
| **Option 3b** | Minimum complexity with over the air signalling; | Optimum | Solved | Feasible | Short |
| **Option 3a/5a** | High complexity in server to server manner; Medium complexity with over the air signalling; | Optimum | Not solved | Infeasible with only short time scale model update | Long |
| **Option 4** | High complexity in server to server manner; Medium complexity with over the air signalling; | Better than Option 1, but worse than Option3 and Option5 | Not solved | Infeasible with short time scale model update | Long |
| **Option 5b** | Need clarification | | | | |

1. **RAN1 concludes that it is recommended to support option 3b to address inter-vendor training collaboration for CSI compression**

OPPO

*Proposal 8:* *Suggest to distinguish the reference model in RAN1 to in RAN4*

* ***Higher requirement on model performance for reference model in RAN1***
* ***RAN1 cannot directly use the agreement on reference model in RAN4***

*Proposal 9: RAN1 further study how to standardize reference model structure in Option 3.*

*Proposal 10: RAN1 further study how to standardize data / dataset format in Option 4.*

*Proposal 11: RAN1 further study how to standardize reference model format in Option 5.*

Xiaomi

***Proposal 1: Parameter or model could be exchanged via standardized signalling, and the following exchange method, format of parameter or model could be considered:***

* ***Model delivery/transfer Case z4 could be used to exchange parameter or model.***
* ***The format could be further studied and defined in 3GPP range if two-sided AI/ML model based CSI compression is studied as a normative work.***

***Proposal 2: Dataset could be exchanged via standardized signalling, and the following definition on format, type and contents of dataset could be considered:***

* ***The format of dataset could be codebook-based quantization (e.g., e-type II like)***
* ***The type of dataset could be eigenvector of raw channel.***
* ***The contents of dataset are target CSI, CSI feedback and/or CSI feedback and reconstructed target CSI.***

***Proposal 3: In order to make the two-sided model be compatible, it is necessary to study the model pairing for Option 3a/3b/4/5a of inter-vendor training collaboration. The pairing information could be assigned by gNB during model/parameter/dataset exchange.***

***Proposal 4: It is necessary to report UE capability on training model, so that gNB could adopt different options for inter-vendor training collaboration.***

***Proposal 5: For Option 3a/4/5a, it is necessary to report the model related aspects, such as scalability, rank and layer handling so that gNB could make suitable parameter configuration for CSI feedback. While it is not necessary to report such model related aspects for Option 3b.***

***Proposal 6: For Option 3a/3b/4/5a, a quantization approach could be standardized reduce the processing complexity of dequantization at gNB or signalling overhead.***

Fujistu

***Proposal 20:***

* *In Option 3a, the model part(s) of the specified reference model structure and the model part(s) of the model parameters exchanged should be the same, i.e.,*
  + *In Option 3a-1, the reference CSI generation part model structure should be specified.*
  + *In Option 3a-2, the reference CSI reconstruction part model structure should be specified.*
  + *In Option 3a-3, both the structures of the reference CSI generation part and the reference CSI reconstruction part should be specified.*

***Proposal 21:***

* *Options 3a-1/5a-1 and Options 3a-3/5a-3 should be deprioritized compared to Options 3a-2/5a-2.*

***Proposal 22:***

* *Option 3a is preferred compared with Option 3b.*

***Proposal 23:***

* *Option 5a is preferred compared with Option 5b.*

***Proposal 24:***

* *RAN1 to further study the content of the dataset(s) of ground-truth CSI for training AI/ML models, which covers, as much as possible, the typical channel conditions of the scenarios of interest.*

***Proposal 25:***

* *If two-sided model for CSI compression will be specified in Rel-19, Option 4 for alleviating/resolving the inter-vendor collaboration issues is recommended for normative work.*

CATT

**Proposal 2: For Option 1, to avoid duplicate work between RAN1 and RAN4, study the aspects different from RAN4 option 3/RAN 4 option 4, and not to study the feasiblility and methodology(s) on fully standardized reference model in RAN1.**

**Proposal 3: If Option 3 is supported, prioritize the solution with model structure of CSI generation part is standardized and model/parameter exchanged from the NW-side to UE-side for CSI generation part.**

Proposal 4: If Option 4 is supported, prioritize the solution with dataset exchanged from the NW-side to UE-side consists of (target CSI, CSI feedback).

Proposal 5: Regarding inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model, deprioritize the solutions with UE-side/NW-side servers involved.

Proposal 6: Regarding inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model, for parameter/dataset/model exchange between UE-side and NW-side(Option 3/4/5), prioritize the solutions with over-the-air signaling standardized, and RRC signaling can be considered as a starting point.

Proposal 7: Down-select options for inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model, and the follow-up study on specification impacts can focus on the selected option(s).

Proposal 8: Regarding inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model, consider the following comparisons of the options:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Inter-vendor collaboration complexity** | **Performance** | **Interoperability and RAN4 testing** | **Feasibility** |
| **Option 1** | Low / None | Limited | Good and up to RAN4  Related to RAN4-Option3 and RAN4-Option4 | Feasible  Large spec effort (initially & for evolution) |
| **Option 2** | Low / None | Limited | Good and up to RAN4  Related to RAN4-Option4 | Feasible  Large spec effort (initially & for evolution) |
| **Option 3a** | Low/none if model parameters transfer/model is exchanged using over-the-air signalling  Medium/high if model parameters /model is exchanged offline | Depends on whether additional information (e.g., dataset or information related to collecting dataset) is provided by NW-side. If the additional information is provided, can be better than Option 1/2, otherwise, can be worse than Options 1/2. | Poor interoperability , Specifying dataset and/or a reference model as in Option 1 maybe needed | Might feasible if dataset or information related to collecting dataset is transferred from the NW-side to the UE-side; otherwise, not feasible  FFS the impacts of the delay from the time the UE/UE-side receives the parameters to the time the UE can apply the model  Large spec effort (initially & for evolution) |
| **Option 3b** | Low/none if model parameters transfer/model is exchanged using over-the-air signalling  Medium/high if model parameters /model is exchanged offline | Better than Options 1/ 2 | Poor interoperability Specifying dataset and/or a reference model as in Option 1 maybe needed | Depends on UE capability, and feasible for UEs capable of updating parameters directly  Large spec effort (initially & for evolution) |
| **Option 4** | Low/none if dataset is exchanged through the over-the-air signaling  Medium/high if dataset is exchanged offline | Impacted by whether the backbone/structure of the CSI generation part applied at UE-side is aligned with the CSI reconstruction part applied at NW-side | Poor interoperability Specifying dataset and/or a reference model as in Option 1 maybe needed | Feasible  FFS the impacts of the delay from the time the UE/UE-side receives the dataset to the time the UE can apply the model.  Low spec effort |
| **Option 5a** | Low/none if model is exchanged using over-the-air signalling  Medium/high if model is exchanged offline | Depends on whether additional information (e.g., dataset or information related to collecting dataset) is provided by NW-side. If the additional information is provided, can be better than Options 1/2/3, otherwise, can be worse than Options 1/2 | Poor interoperability Specifying dataset and/or a reference model as in Option 1 maybe needed | Might feasible if additional information (e.g., dataset or information related to collecting dataset) is transferred from the NW-side to the UE-side; otherwise, not feasible  FFS the impacts of the delay from the time the UE/UE-side receives the model to the time the UE can apply the model  Less spec effort than Options 1/2/3 |
| **Option 5b** | Medium/high | Best | Poor interoperability Specifying dataset and/or a reference model as in Option 1 maybe needed | Depends on UE capability, and feasible for UEs capable of updating parameters directly  Less spec effort than Options 1/2/3 |

Panasonic

**Proposal 1: For Option 1, RAN4 proposed model structure and/or model parameters for the feasibility study of testing options could be starting point.**

**Proposal 2: Option 1 can be used to define some of “minimum performance”. Further extension on top of Option 1 (e.g., the combination with Option 3/4/5) can be considered.**

**Proposal 3: The pros/cons of different options can be summarized as in the following table.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Inter-vendor collaboration** | **Performance** | **Interoperability / RAN4 testing** | **Feasibility** |
| Option 1 | (Concluded in RAN#116bis)  Eliminate the inter-vendor collaboration complexity | Depends on standardized reference model | (Concluded in RAN1#116bis)  Corresponds to RAN4 options, e.g., RAN4-Option 3 or RAN4-Option 4. Further study and final conclusion is up to RAN4.  To identify the cause of the performance degradation is possible | Feasible |
| Option 3a-1 | Low  Parameter exchange (CSI generation part) | Potential to support localized model by updating the parameters.  Less flexible than Option 5. | FFS:  Combination with Option 1 should be considered.  To identify the cause of the performance degradation may be possible | May be feasible |
| Option 3a-2 | Low  Parameter exchange (CSI reconstruction part) | May be feasible |
| Option 3a-3 | Low  Parameter exchange (CSI generation and reconstruction part) | May be feasible |
| Option 3b | Low  Parameter exchange (CSI generation part) | FFS:  Combination with Option 1 should be considered.  It is unclear how to ensure the performance of the two-sided model and how the operator can identify the responsibility if the two-sided model fails in operation in the field. | May be feasible, but more standardization effort than Option 1 may be needed. |
| Option 4-1 | High (if model structure alignment is needed)  Dataset exchange (target CSI, CSI feedback) | Less reliable | FFS:  Combination with Option 1 should be considered.  It is unclear how the UE side can check if the trained CSI generation model is compatible with the CSI reconstruction model at the NW side and whether the paired model fulfil the performance target. | May be feasible, but less reliable |
| Option 4-2 | High (if model structure alignment is needed)  Dataset exchange (CSI feedback, reconstructed target CSI) | May be infeasible |
| Option 4-3 | High (if model structure alignment is needed)  Dataset exchange (target CSI, CSI feedback, reconstructed CSI) | FFS:  Combination with Option 1 should be considered.  To identify the cause of the performance degradation may be possible | May be feasible, but less reliable |
| Option 5a-1 | Medium  Model exchange (CSI generation part) | Potential to support localized model by exchanging the reference model specific to localized area. | FFS:  Combination with Option 1 should be considered.  To identify the cause of the performance degradation may be possible | May be feasible |
| Option 5a-2 | Medium  Model exchange (CSI reconstruction part) | Potential to support localized model by exchanging the reference model specific to localized area. | May be feasible |
| Option 5a-3 | Medium  Model exchange (CSI generation and reconstruction part) | May be feasible |
| Option 5b | High  Model exchange (CSI generation part)  Model structure alignment based on offline collaboration is needed. | FFS:  Combination with Option 1 should be considered.  It is unclear how to ensure the performance of the two-sided model and how the operator can identify the responsibility if the two-sided model fails in operation in the field. | May be feasible, but more standardization effort than Option 1 may be needed. |

TCL

***Proposal 5: Option 1 and 2 should not be standardized, at least they are out of the scope of R19.***

***Proposal 6: RAN 1 should down select among option 3, 4 and 5 considering if unified model format or structure is shared between the NW and UE side model respectively.***

* ***For option 4, there may be no need for offline-engineering.***
* ***Option 3 and 5 are preferred if model structure or format is exchanged from NW to UE.***

LGE

**Proposal #7: Prioritize Option 4 for addressing inter-vendor training collaboration.**

**Proposal #8: Study on model complexity method, e.g., knowledge distillation, to further reduce the CSI training/signaling complexity for Type 3 training collaboration.**

Lenovo

1. Due to performance limitation and also required high specification effort, we suggest deprioritizing Option 1 for inter-vendor training collaboration.
2. For cases with offline engineering, prioritize schemes based on exchange of information regarding the decoder model over options exchanging information of encoder model. It is since, both methods have almost similar 1) signaling overhead, 2) specification complexity, and 3) offline training complexity, while sharing of the decoder model enables the UE to evaluate the performance of the trained encoder model.
3. For options based on exchange of information for both encoder and decoder model, we do not expect better performance compared to cases with exchange of only decoder model. However, information regarding both encoder and decoder can be useful in cases which we intend to implement a UE-sided root-cause procedure. Therefore, if not used for root cause procedure, we suggest deprioritizing Option 3-3, 4-3, and 5-3 in favor of Option 3-2, 4-2, and 5-2, respectively, due to its higher overhead.
4. Prioritize options based on exchange of information regarding the decoder model, i.e., (CSI feedback, reconstructed target CSI) samples (Option 4-2), over exchange of information regarding the encoder model, i.e., (target CSI, CSI feedback) samples (Option 4-1) .
5. Despite potentially much lower complexity, direct use of received parameters (instead of offline engineering) may result in UE encoder with not acceptable performance. Further study is needed in this regard.
6. Until further investigation, give higher priority to options based on offline engineering over options based on direct use of parameters.
7. At least for cases with offline-engineering step, prioritize schemes based on exchange of complete model (or options based on dataset exchange) over options based on exchange of model parameters only.
8. Based on the above observations and proposals, we suggest to prioritize options 1) with offline engineering and 2) are based on transfer of the decoder models, i.e., Options 3a-2, 5a-2 and 4-2. Among these options, to enable possibilities for future enhancement, we propose to further prioritize options 5a-2 and 4-2 (highlighted in the table). In cases where UE-sided root cause procedure is also implemented, we propose to prioritize options 5a-3 and 4-3 (also highlighted in the table).
9. Support definition of pairing information based on the conditions/additional conditions assigned to the samples of the datasets used for training of the model.
10. Further study model identification/selection procedures during inference time when different models have been developed for different UE-NW vendor pairs.

Sony

**Proposal 1: RAN1 should consider option 3 or 5 as baseline for inter-vendor training collaboration of AI/ML-based CSI compression using two-sided models.**

**Proposal 2: RAN1 should support for option 3 or 5 to directly use of the delivered model/parameters at the UE side.**

**Proposal 3: RAN1 should support additional re-training based on provided model/parameters at UE-side.**

**Proposal 4: Option 3a-1/5a-1 or Option 3a-3/5a-3 can be supported for inter-vendor training collaboration. In addition, UE can directly use the provided CSI generation part from gNB to reduce UE-side training complexity.**

Nvidia

**Proposal 5: RAN1 to conclude that it is feasible to resolve issues related to inter-vendor training collaboration for AI/ML-based CSI compression.**

NEC

***Proposal 2:*** ***RAN1 to prioritize the following options for further study to alleviate / resolve the issues related to inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model:***

* ***Option 3: Standardized reference model structure + Parameter exchange between NW-side and UE-side***
* ***Option 5: Standardized model format + Reference model exchange between NW-side and UE-side***

***Proposal 3: For Option 3/5, RAN1 to prioritize the following options:***

* ***Option 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.***
* ***Option 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.***

***Proposal 4: For Option 3a/5a,*** ***RAN1 to prioritize the following options:***

* ***Option 3a-2/5a-2: Model/Parameters exchanged from the NW-side to UE-side is CSI reconstruction part.***

Nokia

Proposal 4: 3GPP needs to downprioritize the flavor of model/parameter/dataset exchange originating from the UE-side and ending at the NW-side as regards inter-vendor training collaboration options.

Proposal 5: In case that downselection can be considered, prioritize Options 4-2, 4-3, which are to be possibly supplemented by Options 3a-2, 3a-3 as regards facilitation of reference CSI reconstruction model training and associated data exchange, whether it be training dataset or trained parameters.

Proposal 6: For Option 4, RAN1 may consult other WGs to identify feasible solutions.

Samsung

**Proposal#14: While evaluating the performance impact of the (sub)options to resolve the inter-vendor training collaboration complexity of AI/ML-based CSI compression, RAN1 to consider the following aspects**

* **Site/cell/location/scenario-specific models**
* **Vendor-specific optimization**

**Proposal#15: For Option 4-1/2/3, data exchange from NW-side to UE-side, consider the following cases for performance**

* **Case 1: UE-side directly trains CSI generation part of two-sided model**
* **Case 2: UE-side trains the CSI generation part after training a nominal CSI reconstruction part of two-sided model**

**Proposal#15: For Option 4-1/2/3, data exchange from NW-side to UE-side, RAN1 to consider whether the target CSI is to be shared before or after quantization**

ETRI

**Observation 3: Regarding options for alleviating inter-vendor training collaborations, followings were observed:**

* **Option based on the dataset delivery (i.e., Option 4) is less feasible, due to the size of datasets and additional training time**
* **Option based on the reference model delivery (i.e., Options 3 and 5) is less feasible, due to the additional training time. Conversely, delivery of the reference model (i.e., Options 3 and 5) for direct use is feasible when it is applicable**
* **Options 3a/5a have high inter-vendor training collaboration complexity due to the offline engineering**
* **Options 3a/5a, datasets are additionally required for performance assessment**
* **Options 3a/5a may have reduced alignment performance compared to Options 3b/5b**
* **Option 3 is expected to require significantly larger standardization efforts compared to Option 5**
* **Option 3 may have performance constraints compared to Option 5, as it requires the application of a standardized reference model structure, limiting the flexibility of the model architecture.**

MTK

1. Discuss how option 1 and option 3 will ease inter-operability test efforts and also candidate scenarios/config as starting point.

Apple

**Proposal 1: Support option 3a-2, 3a-3, option 5a-2, 5a-3, and option 4-2 and option 4-3 which provide information related to CSI reconstruction model. Reference CSI reconstruction model can ensure E2E performance target, handle UE side additional condition and enable extendibility of future releases.**

Qualcomm Incorporated

1. Recommend option 3a-2 / 5a-2 with CSI reconstruction part sharing, or option 4-2 via sharing the input / output of CSI reconstruction part.

* For CSI reconstruction part sharing, the signalling can be either over-the-air-interface, or other signalling to be determined by other working group.
* For dataset sharing 4-2, recommend approaches other than over-the-air-interface signalling.

1. Specification of model LCM aspects should accommodate models that are designed via proprietary structure and proprietary signallings.

CAICT

***Proposal 1: Option 3a/3b/4 should be considered to be specified for two-sided model inter-vendor training collaboration.***

***Proposal 2: Option 5a/5b could be deprioritized.***

NTT Docomo

**Proposal 1**

* **The conclusion on the interoperability/testability of inter-vendor collaboration options is up to the RAN4.**

**Proposal 2**

* **RAN1 focuses on the study of Option 1 and Option 3a and concludes the feasibility of Option 1 and Option 3a based on the outcome of RAN4.**

CEWiT

**Proposal-3: In case of standardized model, down-selection on which part of the model is to be needed. The decoder is expected to be started with for model standardization.**

**Proposal-4: In terms of Option 1/3 for specifying model structure, consider the following aspects:**

* **Model architecture**
* **Number of layers**
* **Hyper-parameters**
* **Quantization method**

**Proposal-6: The model parameter exchange between the UE side and the NW side should be specified after the model identification and selection process.**

**Proposal-9: In case of improving inter-vendor collaboration, store the additional information of an NW-sided model like vector-quantisation codebook name or its properties (size, feature length).**

**Proposal-10: In case of Type-III UE first raining, train the CSI reconstruction model with the knowledge of UE specific codebook.**

ITL

***Proposal 1: If RAN4 specifies the reference model that well is aligned RAN1 Option 1, the option 1 in RAN1 can be supported with moderate specification efforts to enhance the AI/ML model performance based on the reference model.***

***Proposal 2: Option 3a-1 and 3a-3 can be prioritized than Option 3a-2***

***Proposal 3: For option 3b, it should be clarified what model parameters/input data processing needs to be standardized together with the CSI generation model structure. Otherwise, Option 3b is less feasible than Option 3a when considering there are many multiple vendors.***

***Proposal 4: It is proposed that Option 5a is deprioritized for inter-vendor training collaboration.***

***Proposal 5: The standalone use of Option 4 is not preferred. It is recommended to use it as a supplementary option alongside other options.***

## Discussion

|  |  |
| --- | --- |
| Companies | Views |
| Futurewei | Upper layer signaling as delivery option for 3a/5a/4  Quantization for inter-vendor issue for option 3a/5a/5/3b: VQ/SQ, configuration, exchange of codebook |
| Huawei | Capture comparison table among options  RAN1 need more time / effort to justify standardization feasibility   * Option 1/3: model aspects * Option 4/5 dataset format, model format * Option 3/4/5: signaling * Downselection among options |
| Spreadtrum, BUPT | Further study dataset format, model format / structure |
| Google | Prioritize option 3/5, whether UE use it directly is upto implementation |
| Tejas Networks | Prioritize option 3 over 5,  3b over-the-air, 5b offline signalling  Option 4 needs offline mechanism for dataset exchange |
| CMCC | Progress in RAN4 can reused for option 3 study  Model which is exchanged aligned with model whose structure is specified  Discuss feasibility, performance, complexity together  At least 3a, 5a, 3b prioritized |
| Intel | 3b is priorized over 3a  UE not expected to satisfy the performance target provided  NW provide guidance on CSI quality, or UE report CSI quality for NW to further utilize for LCM  Mapping options to training collaboration types |
| ZTE | Option1 consider study of encoder  Further study option suboptions performance  Prioritize OTA signaling for option 3/4/5  Post development operation during actual deployment to guarantee interoperability  Further study unified solution for model identification, multi-vendor collaboration, model pairing |
| Ericsson | On top of RAN1 options, UE/NW side can further improve their model using dataset collected in the field.  Option 5a, 3a-2/3, 4-2/4-3 not supported  Option 3a-1 and 4-1 go through offline signalling  For option 3b, needs to standardize parameter precision, pre-processing, additional information, model structure. Need to study their feasibility of standardization and complexity  Option 4, additional information of pre-processing, post processing, training objectives is needed |
| vivo | Procedures of specifying model structure   * Aligning EVM, determine backbone, hyper-parameters   Recommend option 3b |
| OPPO | Ran1 cannot directly use the agreed ref model from RAN4, ran1 has high requirement on model performance  Further study how to standardize model structure, dataset format, model representation format |
| Xiaomi | Z4 could be used for model / parameter exchange, format could be studied within 3gpp  Dataset format includes, Et2 like precoder or channel. Contents include {target CSI, CSI feedback} or {reconstructed CSI, CSI feedback}  Other aspects: model pairing, UE capability, model related aspects (scalability, layer handling) and standardized quantization |
| Fujistu | Prioritize 3a-2 over 3a-1, 3a-3, 3b  Prioritize 5a over 5b  Recommend option4, further study content of dataset |
| CATT | For option1, study different aspects than RAN4  Prioritize option 3a-1, 3b over 3a-2, 3a-3  Prioritize 4-1 over 4-2, 4-3  Deprioritize UE/NW side server involved  Prioritize over-the-air signalling  Down-select among suboptions |
| Panasonic | RAN4 proposed model can be starting point for RAN1 option 1  Option 1 can be used for minimum performance |
| TCL | Option 3 and 5 are preferred, need further downselection |
| LGE | Prioritize option 4, study model complexity method to further reduce signalling complexity |
| Lenovo | Deprioritize option 1  Prioritize 4-2 over 4-1, 4-3  Prioritize 3a-2, 5a-2 over 3a-1/3, 5a-1/3   * If UE side root cause identification, prefer 3a/5a-3   Option 3b/5b may have unacceptable performance, needs further study  Prioritize option 5 over 3, offline engineering over on-device operation  Further study model identification, pairing, selection procedure |
| Sony | Prioritize option 3a/5a-1 or 3a/5a-3 |
| Nvidia | RAN1 conclude it is feasible to resolve inter-vendor collaboration issue |
| NEC | Prioritize 3a/5a-2 |
| Nokia | Prioritize option 4a-2/3 or 3a-2/3.  RAN1 consult other WGs for feasible signaling option |
| Samsung | While evaluating performance of sub-options, RAN1 further consider localized models and vendor-specific optimization  For option 4, further consider whether UE directly trains encoder, or train a decoder first and followed by actual encoder  For option 4, consider dataset shared before or after quantization |
| MTK | Study how option 1 and 3 can ease inter-vendor efforts |
| Apple | Support 3a/5a/2-/3, 4-2/3 |
| QC | Recommend 3a/5a-2, 4-2.   * 3a/5a-2 with OTA or offline signaling, 4-2 with offline signaling   Proprietary model and signalling should be accommodated by spec. |
| CAICT | Deprioritize option 5a and 5b |
| DCM | Focuses on option 1 and 3a |
| ITL | Option 1 efforts is moderate considering RAN4 outcome  Option 3a-1/3 preferred over option 3a-2  Option 5a deprioritized, option 4 not used standalone  Option 3b needs further clarification on parameter/input data processing. |

|  |  |  |
| --- | --- | --- |
| Aspects | comment | Companies |
| **Down-selection among sub-options** | **3a-1 (or with 3a-3)**   * UE implementation of either on-device operation or offline engineering * No disclosure of NW proprietary information | (8) Google, Tejas Networks, CMCC, Ericsson, CATT, Sony, DCM, ITL |
| **3a-2 (or with 3a-3)**   * Better performance * easy for UE side monitoring | (9) Tejas Networks, CMCC, Fujistu, Lenovo, NEC, Nokia, Apple, QC, DCM |
| **3b**   * Easy UE implementation, shorter deployment time | (6) Google, Tejas Networks, CMCC, Intel, Ericsson, vivo, CATT |
| **Deprioritize option 5**   * considering inter-vendor complexity aligning model structure | (3) Ericsson, CAICT, ITL |
| **Supportive of option 5** | (7) Google, CMCC, Fujistu, Lenovo, Sony, Apple, QC |
| **4-1 (or with 4-3)**   * With similar consideration as for 3a-1 | (4) Ericsson, Fujistu, CATT, LGE |
| **4-2 (or with 4-3)**   * With similar consideration as for 3a-2 | (6) Fujistu, LGE, Lenovo, Nokia, Apple, QC |
| **Need further study** | (3) Huawei, ZTE, TCL |
| **Signaling** | Upper-layer (offline) signaling | (5) Futurewei, Tejas Network (5b), Ericsson, Nokia, QC |
| OTA signaling | (4) ZTE, Tejas Network (3b), CATT, QC (if option 3a-2) |
| Further study | Huawei |
| **Option 1 aspects** | RAN1 can leverage progress of RAN4 while considering higher performance requirement | OPPO, CATT, Panasonic |
| UE/NW side can optimize their model using field data based on reference model in option 1. | Ericsson |
| Option 1 considered as the minimum performance | Panasonic |
| Specify encoder in option 1 | ZTE |
| **Further study aspects** | Feasibility of model structure and related aspects (option 3) | Huawei, vivo, OPPO, Xiaomi |
| Dataset format and details (option 4) | Huawei, Spreadtrum, OPPO, Xiaomi |
| Model representation format (option 5) | Spreadtrum, OPPO, Xiaomi |
| Additional information, parameter precision, pre/post processing (option 3/4/5) | Ericsson |
| Model identification, pairing, quantization alignment, etc | Futurewei, ZTE, Xiaomi, Lenovo, Samsung |

Here is a quick summary of status:

Option 1

* Conclusion: Interoperability
* FFS: limited performance in the field, feasibility of specification, necessity of being used as baseline, or RAN4 itself is used as baseline

Option 3

* Inter-vendor collaboration alleviated/resolved by standardized signalling, feasibility of model structure specification proved by RAN4, performance monitoring to ensure field performance.
* FFS:
  + Sub-option comparison considering data distribution mismatch, proprietary information disclosure, comparison of additional information to be shared.
  + Feasibility of model structure specification
  + Feasibility of specified signaling - OTA vs. specified offline signaling (consult other WG) considering feasibility and overhead.
  + 3b device capability

Option 4:

* Inter-vendor collaboration alleviated/resolved by standardized signalling, performance monitoring to ensure field performance
* FFS:
  + Sub-option comparison considering data distribution mismatch, proprietary information disclosure, comparison of additional information to be shared.
  + Feasibility of specifying dataset sharing signalling and content - OTA vs. specified offline signaling (consult other WG) considering feasibility and overhead.

Option 5:

* Option 5 alone doesn’t fully address inter-vendor collaboration complexity but provides potential performance benefit and flexibility on top of Option 3 for vendors willing to do inter-vendor collaboration. Option 5 can be easily supported if Option 3 is supported. Study feasibility and additional specification impact of supporting Option 5 in comparison to Option 3

### Down-selection of options

There is competitive supporting of option 3a/5a-1, 3a/5a-2, 3b, 4-1, 4-2

* Main difference between 3a/5a-1 (4-1) vs. 3a/5a-2 (4-2): encoder (info) or decoder (info) sharing
* Main difference between 3a/5a-1 vs. 3b: same for encoder sharing, UE offline engineering can be left for UE implementation, hence the main difference lies in signaling, deployment mechanism.

Proposal 21a:

Confirm the necessity of both directions - on-device operation and UE side offline engineering - and recommend both of them for normative work, contingent on their feasibility

* Direction A: Sharing reference model / parameters / dataset that enables UE-side offline engineering (3a and 4)
  + FFS: down-selection into one or more among sub-options considering their feasibility and performance
  + Note: Option 5a is discussed along with 3a with the understanding that model structure is aligned offline among vendors.
  + Note: Option 3a with and without dataset sharing can be considered
  + Note: For performance study, companies are encouraged to provide evaluations considering mismatch between NW side data distribution and UE side data distribution due to UE-side additional condition and pre-processing, e.g., hardware impairments, SVD algorithm, antenna imbalance, etc.
    - Encoder vs. decoder exchange
    - Dataset vs. model / parameter exchange
* Direction B: Sharing NW side encoder parameter to UE side for UE side inference directly with on-device operation (3b)

|  |  |
| --- | --- |
| *Support / Can accept* |  |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Support of this direction. |
| Xiaomi | Support to study this direction |
| Huawei, HiSilicon | 1) For the recommendation to normative work, we think it is premature. Reasons:  i) The feasibility of these options is not justified. E.g., whether it is feasible to converge on specified and optimized model structure for 3a/3b, what is the model representation format for 5a, whether the signalling is over the air or other approaches (and what is the other approach). Before we justify the feasibility, we may not able to do the recommendation.  ii) The evaluations/spec impact analysis/down selection of new cases (see proposal 2a/3a) are not completed.  iii) Whether it is feasible to enable NW side data collection and NW side monitoring based on high resolution ground-truth CSI, and UE side monitoring based on proxy model (or other approaches).  It is premature to recommend normative work before justifying feasibility for these aspects. Therefore, we suggest a continued study of both directions.  2) If the assumption of Option 5a is offline alignment across vendors, then Option 5a can be deprioritized, since it does not alleviate cross vendor collaboration.  3) To alleviate the study effort, we can focus on NW-first and NW-side training. So, a note is added.  **Suggested changes:**  ~~Confirm the necessity of both directions - on-device operation and UE side offline engineering - and recommend both of them for normative work, contingent~~ Continue the study for both directions on their feasibility   * Direction A: Sharing reference ~~model /~~ parameters / dataset that enables UE-side offline engineering (3a and 4)   + FFS: down-selection into one or more among sub-options considering their feasibility and performance   + ~~Note: Option 5a is discussed along with 3a with the understanding that model structure is aligned offline among vendors.~~   + Note: Option 3a with and without dataset sharing can be considered   + Note: For performance study, companies are encouraged to provide evaluations considering mismatch between NW side data distribution and UE side data distribution due to UE-side additional condition and pre-processing, e.g., hardware impairments, SVD algorithm, antenna imbalance, etc.     - Encoder vs. decoder exchange     - Dataset vs. model / parameter exchange * Direction B: Sharing NW side encoder parameter to UE side for UE side inference directly with on-device operation (3b) * Note: the study focuses on NW-first training and NW-side training. |
| Lenovo | We support the direction of the proposal. The following “Note” is not clear to us:  Note: Option 3a with and without dataset sharing can be considered |
| NEC | Support to study the two directions. |
| Panasonic | We are fine with the proposed direction. |
| Futurewei | Originally these 5 options were identified for the discussion of alleviating/resolving inter-vendor training collaboration issue, thus, we agree this aspect is to be considered and the first part of this proposal is from this angle. However, “recommend both of them for normative work, contingent on their feasibility” is controversial, as we should at least have consensus on feasibility first before recommending them for normative work. |
| ETRI | Support to consider the two directions. |
| SK telecom | Support to prioritize these two directions than others |
| LG | Fine with the proposal |
| TCL | Agree |
| OPPO | Support to study this direction. |
| ZTE | For “recommend both of them for normative work”, we also think it is not mature enough. Reasons are listed as below:  **The feasibility issue of all the options are not discussed and justified yet.**   * The feasibility issue is hard to be determined only in RAN1 working group, since the interoperability/RAN4 testing issue needs to be studied and discussed in RAN4 group. The feasibility verification can be further decided based on RAN1 and RAN4 concrete outcomes. * From the perspective of performance, evaluations are not performed by companies to verify the feasibility of different options. * The spec impact analysis is not further discussed in detail.   So, we think more study is needed on the both directions.  For the main bullet, we suggest changing it as  ~~“Confirm the necessity of both directions - on-device operation and UE side offline engineering - and recommend both of them for normative work, contingent~~ Further study the feasibility of both directions”  For the 3rd bullet of Direction 1, we are not clear whether only dataset sharing is mentioned here, which needs a clarification. According to the former agreement, to our understanding, dataset sharing is one factor in additional information (e.g., performance target other than dataset) if necessary, so we think it should also be additional information to be aligned with the previous agreement. So, we suggest changing it as  Note: Option 3a with and without ~~dataset sharing~~ additional information can be considered |
| Ericsson | We agree with Huawei that it is premature to recommend normative work, since the feasibility of the inter-vendor training option 3/4/5 have not been concluded. In addition, companies have very different views on which options shall be down selected. The study should be down selected to one option since we shouldn’t standardize several options for the same functionality. Work item description should have very clear objective on what should be specified rather than discussion down-selection of many different options.  We also with Huawei that option 5a requires significant inter-vendor collaboration, hence, should not be discussed in 3GPP at all.  Please find below further changes in green based on Huawei’s version:  Continue the study for both directions on their feasibility   * Direction A: Sharing reference ~~model /~~ parameters / dataset that enables UE-side offline engineering (3a and 4)   + FFS: down-selection into one or more among sub-options considering their feasibility and performance   + Conclude on the necessary additional information to be exchanged from the NW-side to the UE-side to help UE-side offline engineering, model training and performance validation.   + ~~Note: Option 5a is discussed along with 3a with the understanding that model structure is aligned offline among vendors.~~   + Note: Option 3a with and without dataset sharing can be considered   + Note: For performance study, companies are encouraged to provide evaluations considering mismatch between NW side data distribution and UE side data distribution due to UE-side additional condition and pre-processing, e.g., hardware impairments, SVD algorithm, antenna imbalance, etc.     - Encoder vs. decoder exchange     - Dataset vs. model ~~/~~ parameter exchange * Direction B: Sharing NW side encoder parameter to UE ~~side~~ for UE ~~side~~ inference directly with on-device operation (3b)   + Study the feasibility and necessity of standardizing the model parameter precision and input data pre-processing for the encoder.   + Conclude on if there is any other additional information that needs to be standardized to improve the feasibility for a UE to use received parameters directly for inference.   Note: the study focuses on NW-first training and NW-side training.  Note: for both directions, the feasibility of performance testing/monitoring with low complexity of inter-vendor offline collaboration shall also be considered when drawing conclusions. |
| Spreadtrum | Support to study this direction. And we want to make sure the proposal means option 5 is deprioritized, right? |

### Model structure standardization (for Option 1 and 3)

Regarding model structure and model standardization in option 3 and option 1,

* Some companies propose that RAN4 procedure can be leveraged
* Some companies comment that feasibility needs to be studied before recommending normative work.
* Sne company proposes suggested procedure, e.g., aligning EVM, backbone, hyper-parameter, model scalability / adaptation.

Proposal 22a:

For model and model structure specification in option 1 and option 3, conclude that

* Option 1 is feasible if RAN4 confirms the feasibility of specifying an encoder or decoder and its parameters.
* Option 3a is feasible if RAN4 confirms the feasibility of specifying an encoder or decoder.
* Option 3b is feasible if RAN4 confirms the feasibility of specifying an encoder or decoder, pending feasibility of UE implementation feasibility of the encoder.

|  |  |
| --- | --- |
| *Support / Can accept* |  |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Supportive of the direction.  For RAN4 to confirm the feasibility, not sure whether RAN4 would conduct such feasibility confirmation and whether such confirmation would block RAN1 progress. Thus we change it to a more general wording as “if RAN4 discussion show the feasibility to align on”:  For 3b, we did not fully understand the original intention. If it is talking about the UE capability for Option 3b, we would like to reformulate the wording as following:  Proposal 22a:  For model and model structure specification in option 1 and option 3, conclude that   * Option 1 is feasible if RAN4 discussion show the feasibility to align on ~~confirms the feasibility of specifying an~~ model structure of encoder or decoder and its parameters. * Option 3a is feasible if RAN4 discussion show the feasibility to align on ~~confirms the feasibility of specifying an~~ model structure of encoder or decoder. * Option 3b is feasible if RAN4 discussion show the feasibility to align on ~~confirms the feasibility of specifying an~~ model structure of encoder or decoder~~, pending feasibility of UE implementation feasibility of the encoder~~. * UE capability for Option 1/3a/3b can be separately specified for each option. |
| Xiaomi | Support |
| Huawei, HiSilicon | 1) For Option 1/3, disagree with reusing RAN4 outcome. RAN4 only targets to specify the minimum requirement that can support passing the testing. But RAN1 should target to specify the optimized model that is running in field. In particular:  i) RAN4 has decided to adopt CNN as the backbone, yet this model structure is much less performed than TF backbone from RAN1 evaluations in R18.  ii) RAN4 only studies the SF domain only CSI compression in R19 up until now, yet RAN1 targets to have advanced model structure to involve temporal domain CSI compression.  iii) As far as we know, RAN4 does not consider scalability over Tx ports, CSI feedback payload sizes, bandwidths. But for field usage, the scalability needs to be considered by RAN1.  2) For Option 3a, this proposal only focuses on the model structure specification, yet the feasibility of signaling the model/parameter with low complexity of cross-vendor collaboration is a separate discussion – e.g., if it is transferred via offline, the cross-vendor collaboration complexity is not alleviated.  **Suggested changes:**  For model and model structure specification in option 1 and option 3, conclude that   * The feasibility of Option 1 ~~is feasible~~ depends on if RAN1 ~~RAN4~~ confirms the feasibility of specifying an encoder or decoder and its parameters. * The feasibility of Option 3a ~~is feasible~~ depends on if RAN1 ~~RAN4~~ confirms the feasibility of specifying an encoder or decoder.   + Note the feasibility of signaling the parameters with low complexity of cross-vendor collaboration is a separate discussion * The feasibility of Option 3b ~~is feasible~~ depends on if RAN1 ~~RAN4~~ confirms the feasibility of specifying an encoder ~~or decoder~~, pending feasibility of UE implementation feasibility of the encoder. |
| Lenovo | For option-1 we believe RAN4 may decide to specifying an encoder or decoder for testing requirement, but as we discussed in RAN1, option-1 has limitations that may not be considered for solving the interoperability issues.  On the other side, option 3a/3b may be used only for interoperability issues in RAN1 and even not discussed in RAN4. |
| Panasonic | We are fine with the proposal. |
| LG | Similar view with Huawei that the options listed above can meet not only for the testability but also at least for the aspects on inter-vendor training collaboration. |
| TCL | Agree |
| OPPO | We support to start standardization of model and model structure in RAN1. But the performance requirement for RAN1 is much higher than RAN4. So RAN1 cannot directly reuse the specified reference encoder and decoder in RAN4. |
| ZTE | We are fine with the proposal. |
| Ericsson | For option 1, suggest making the following changes   * The feasibility of Option 1 ~~is feasible~~ depends on if RAN4 confirms the feasibility of specifying a reference/test encoder or reference/test decoder and its parameters.   Our view is that if RAN4 test cannot reflect reality in a meaningful way, then defining and performing a RAN4 test doesn’t have any purpose. Hence, RAN4 shall consider the aspects of temporal domain compression, the scalability over Tx ports, CSI feedback payload sizes, bandwidth, number of ranks/layers, etc., when concluding the feasibility of specifying a reference/test encoder or reference/test decoder for a selected CSI compression use case.  For option 3a, the feasibility of standardized signalling for NW-side to exchange model parameters together with any other necessary information to the UE-side also needs to be concluded:   * The feasibility of Option 3a ~~is feasible~~ depends on if RAN4 confirms the feasibility of specifying a reference/test encoder or reference/test decoder.   + Note the feasibility of signaling the model parameters and any other additional necessary information from NW-side to UE-side with low complexity of cross-vendor collaboration is a separate discussion.   For option 3b,   * The feasibility of Option 3b ~~is feasible~~ depends on if RAN4 confirms the feasibility of specifying a reference/test encoder ~~or decoder~~, pending feasibility of UE implementation feasibility of the encoder. |
| Spreadtrum | Fine with the proposal. |

### Model structure for standardization (for Option 1 and 3)

FL observation:

For Case 0, among precoding matrix (i.e., eigenvectors) and raw channel matrix, majority of companies used precoding matrix in their evaluations. Thus, it is proposed to use precoding matrix in the standardized structure discussion for RAN1’s model structure standardization for inter-vendor collaboration Option 1&3.

Furthermore, for higher rank handling, majority of companies used

* Option 2-1: layer specific and rank common (different models applied for different layers; for a specific layer, the same model is applied for all rank values)
* Option 3-1: layer common and rank common (A unified AI/ML model is applied for each layer under any rank value to perform individual inference)

Thus, it is proposed to use a model structure that operates on a per-layer basis

Several companies noted the complexity benefit of using transformed domain (i.e., angular-delay coefficient). Therefore, such an approach can be considered for further discussion on model structure standardization.

For Case 2 and 3, many companies reused the backbone structure of Case 0, by adding some layers or operations either at the target CSI domain (i.e., before the Case 0 CSI generation backbone and after the Case 0 CSI reconstruction backbone) or at the latent domain (i.e., after the Case 0 CSI generation backbone and before the Case 0 CSI reconstruction backbone). Examples of the former is LSTM or Conv-LSTM over multiple CSI instances. Examples of the latter is differential or recurrent quantizer applied over compressed latent vectors. These approaches are beneficial – both in terms of model complexity, specification effort, and potential seamless operation between Case 0 and 2/3 – in that they reuse the Case 0 model structure toward Case 2 and Case 3.

Proposal 23a:

For model structure standardization in RAN1, RAN1 to prioritize model structures based on the following assumptions:

* Precoding matrix as an input (as opposed to raw channel matrix)
  + Per-layer structure (corresponding to Option 2 and Option 3 for handling rank ≥ 1)
* For temporal domain aspects Case 2 and Case 3, strive to reuse the model structure of Case 0, with additional layers or operations either at the input/output domain or at the latent domain.
* For Case 0, use spatial-frequency domain as a baseline. Angular and/or delay domain representation such as eType-II W2 may also be considered.
* For Case 2 and Case 3, use spatial-frequency domain as a baseline for each CSI observation instance. Angular, delay, and Doppler domain representation such as eType-II W2 may also be considered.

|  |  |
| --- | --- |
| *Support / Can accept* |  |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Support |
| Xiaomi | Support to study this direction. But, for the second bullet, the model struction of Case 2 and Case 3 may be different from that of Case 0. In general, the model struction of Case 2 and Case 3 is much simpler than that of Case 0. It is not necessary to restrict to reuse the model structure of Case 0 for Case 2 and Case 3. |
| Huawei, HiSilicon | If the intension of this proposal is to align the assumption for specifying model structure, then we have two comments:  1) Should we also align the backbone (such as TF)?  2) Should consider scalability for usage in field.  **Suggested changes:**  For model structure standardization in RAN1, RAN1 to prioritize model structures based on the following assumptions:   * Precoding matrix as an input (as opposed to raw channel matrix)   + Per-layer structure (corresponding to Option 2 and Option 3 for handling rank ≥ 1) * For temporal domain aspects Case 2 and Case 3, strive to reuse the model structure of Case 0, with additional layers or operations either at the input/output domain or at the latent domain. * For Case 0, use spatial-frequency domain precoding matrix as a baseline. Angular and/or delay domain representation such as eType-II W2 may also be considered. * For Case 2 and Case 3, use spatial-frequency domain precoding matrix as a baseline for each CSI observation instance. Angular, delay, and Doppler domain representation such as eType-II W2 may also be considered. * For Case 0, 2, and 3, use Transformer as the backbone. * Consider scalability approaches over numbers of Tx ports, CSI feedback payload sizes, and bandwidths. |
| Lenovo | We support the proposal and also support HW changes just without the point on decision on the use of Transformer as the backbone (this part may be postponed to a later stage) |
| Panasonic | For temporal domain aspects Case 2, if per-layer structure is used, rank adaptation handling should be studied. Rank-common AI/ML model setting has possibility to simply solve the issue of rank adaptation. |
| Intel | Fine with the proposal. |
| LG | Fine with the direction. Based on per-layer structure, it would be possible to handle the rank adaptation using the past CSI for each layer. |
| TCL | Agree |
| OPPO | Support in general. And Case 0 with only SF-domain CSI compression should be treated in higher priority. Model with temporal aspects can be discussed after the model for Case 0 is stable. |
| ZTE | For the 2nd bullet, we have a question whether the model structure is limited to TF? since some companies may not use the TF for Case 3. |
| Ericsson | We would like to understand the intention for this proposal, why such study shall be conducted in RAN1, considering that RAN4 will anyway standardize reference/test encoder/decoder for use case 0/2/3, if any of these use case is selected. RAN1 and RAN4 shouldn’t do double work.  Is the intention to let RAN1 study the reference/test encoder or reference/test decoder, and then RAN1 provides the study output to RAN4, so that RAN4 implements the reference/test encoder or reference/test decoder in RAN4 specifications? |

### Model parameter / dataset exchange

Question 24a:

* How to confirm the feasibility of dataset exchange, e.g., study the overhead? Can we confirm it is feasible if NW-side data collection is feasible?
* How to confirm the feasibility of model parameter exchange in option 3 / 5?

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | Need to confirm the feasibility of the signalling of dataset/model/parameters. E.g., for over the air signalling, whether it is possible to alleviate the power consumption/overhead with some solutions? This study is commonly applicable to option 3/4/5, and may possibly involve RAN2 (E.g., CP? UP? Data/model segmentation?). |
| Ericsson | For option 3/4/5, we need to first conclude on what are the necessary information that needs to be exchanged from NW side to UE-side for assisting UE-sided offline engineering, model training, performance validation. After that, we need to conclude on the feasibility of different signalling methods for exchanging the information.  For information exchange signalling method, at least the following aspects shall be considered when study its feasibility: inter-vendor offline collaboration complexity, signalling overhead, signalling complexity, UE power consumption, impact on network’s normal services, latency (short model-in-use time). |

### Others

Please provide any other comments for this section.

|  |  |
| --- | --- |
| *Company* | *Comments* |
|  |  |
|  |  |

# Data collection

## Summary of company proposals

From the submitted contributions, proposals related to data collection are summarized below.

Futurewei

***Proposal 3: In AI/ML-based CSI compression using two-sided model, at least for NW-side ground-truth data collection for model training, consider adopting Rel-16 eType II CB based quantization with new parameters to achieve better performance.***

Huawei

***Proposal 6: For the NW side data collection, confirm the necessity and feasibility of UE report of the ground-truth CSI.***

* ***For the data sample type, prioritize precoding matrix over channel matrix.***
* ***For the data sample format, prioritize Rel-16 eType II CB based quantization with new parameters, and take the following new parameters (captured in the Rel-18 observation) as candidates for discussion.***
  + ***L= 8, 10, 12; pv = 0.8, 0.9, 0.95; reference amplitude = 6 bits, 8 bits; differential amplitude = 4bits; phase = 5 bits, 6 bits.***
* ***For the number/index(es) of layers*** ***for the collected ground-truth CSI, it can be indicated by NW.***
* ***To alleviate UE complexity, it can be considered to limit the number of subband for each CSI report.***

***Proposal 7: In CSI compression with training collaboration Type 3, the following aspects could be further studied for over-the-air dataset delivery from RAN1 perspective, including:***

* ***Dataset ID, which is used to differentiate the models to be trained at the opposite side.***
* ***Dataset size, e.g., the number of data samples contained in the delivered dataset.***

***Proposal 8: For the dataset delivery of CSI compression over air-interface, study the solution to relieve the overhead.***

* ***E.g., NW splits the overall dataset into many subsets each with a limited number of data samples (e.g., with an overhead comparable to the RRC signaling). The subsets can be separately sent to different UEs, and all subsets are associated with a common dataset ID for the UE side re-combination.***

Google

***Proposal 7: Support to configure the number of layers for the report for NW side data collection for performance monitoring.***

***Proposal 8: Support to report singular values for the ground-truth CSI.***

***Proposal 9: Support to report CQI/RI in addition to the ground-truth CSI.***

***Proposal 10: Reuse the existing CPU framework to handle the UE complexity for the measurement and report for NW side data collection.***

***Proposal 11: Support to maintain the same understanding between the NW and UE on when to perform the measurement for UE side data collection based on the following options:***

* ***Option 1: The measurement for UE side data collection is configured by the NW***
* ***Option 2: UE request CSI-RS for data collection***

Tejas Networks

**Proposal 12: For the NW side data collection prioritize UE reporting precoding matrix than UE reporting channel matrix as a ground truth.**

CMCC

***Proposal 11: In CSI compression using two-sided model use case, regarding the ground truth CSI format for NW side data collection for performance monitoring and model training, R16 eType****-****II codebook and Rel-18 Doppler codebook can be used as a starting point.***

***Proposal 12: In CSI compression using two-sided model use case, regarding the ground truth CSI format for NW side data collection, the basic codebook structure could be reused, along with the basic concept of spatial domain, frequency domain and Doppler domain basis.***

***Proposal 13: In CSI compression using two-sided model use case, regarding the ground truth CSI format for NW side data collection, the exact supported values of codebook parameters can be studied to make sure high resolution data report.***

ZTE

***Proposal 15:*** *For network side data collection, support to further study*

* *Enhanced Rel-16 eTypeII codebook design to achieve high-resolution CSI for model training and performance monitoring*

***Proposal 16:*** *To enable high-quality data collection from UE to network, at least support*

* *UE reports data quality related information to NW,* *e.g., SINR, CQI, positioning information*
* *NW configures a threshold of data quality to UE and UE only reports the qualified data to NW*

OPPO

*Proposal 12: Regarding the CSI-RS configuration for UE-side and NW-side data collection, at least the follow aspects should be considered:*

* *Data collection specific CSI-RS or not*
* *Cell-specific or UE specific CSI-RS*
* *Trade-off between performance and overhead*

*Proposal 13: Regarding the data collection for CSI compression, cell/site/scenario related “condition information” and “addition condition information” should be considered during the data collection stage*

* *Condition information including CSI-related information such as the CSI type, e.g. raw channel or precoding matrix, and the CSI configurations, e.g. number of antenna ports, number of sub-bands, ranks.*
* *Additional condition information including cell/site/scenario related information such as cell/site/scenario ID, indoor/outdoor indication, LoS/NLoS flag and UE ID.*

Fujistu

***Proposal 19:***

* *In AI/ML-based CSI compression using two-sided model use case, the spatial-layer-related information should be included in the report of ground-truth CSI by higher-resolution e-type II like codebook for NW-side performance monitoring and data collection for training.*

CATT

**Proposal 9: In CSI compression using two-sided model use case, for NW-side data collection for model training, focus on CSI-RS measurement based data collection.**

**Proposal 10: In CSI compression using two-sided model use case, both L1 signalling based reporting and RRC signalling based reporting are supported for ground-truth CSI for NW-side data collection for model training.**

**Proposal 11: In CSI compression using two-sided model use case, L1 signalling based reporting of target CSI is supported for NW-side data collection for performance monitoring.**

**Proposal 12: In CSI compression using two-sided model use case, for L1 signalling based reporting of ground-truth CSI/target CSI for NW-side data collection, legacy CSI feedback framework can be reused for Case 0.**

**Proposal 13: In CSI compression using two-sided model use case, for L1 signalling based reporting of ground-truth CSI/target CSI for NW-side data collection, study whether multiple CSI in the same report is supported for Case 2/3/4.**

**Proposal 14: In CSI compression using two-sided model use case, for NW-side data collection for model training, collecting ground-truth data in type of precoding matrix is supported.**

**Proposal 15: In CSI compression using two-sided model use case, for NW-side data collection for model training, NW determines the number of layers for ground-truth CSI data collection.**

Panasonic

**Observation 14: Data collection for model training and non-real time (slow) monitoring is not required to be real-time and then latency requirement can be relaxed.**

**Observation 15: Ground-truth CSI reporting could be realized through U-plane at least for data collection for model training and non-real time (slow) monitoring.**

**Observation 16: Assuming fast monitoring is 100s of ms order, U-plane, RRC or MAC-CE can be sufficient.**

**Observation 17: On data sample type / format for ground-truth CSI reporting, high resolution codebook-based format e.g., legacy codebook (e.g., eType II codebook) with potential enhancements such as extend more configurations in some parameters, should be studied.**

**Observation 18: For NW-side data collection, at least time stamps / situation of measurement, cell ID and UE location should be considered as the UE-side additional condition.**

**Observation 19: For NW-side data collection, the necessity and feasibility of UE reporting Rx filter assumption to network should be studied. Instead of informing actual configuration, UE-side associated ID is necessary.**

**Observation 20: For UE-side data collection for UE side training, in order to identify the scenario / configuration, how to share the NW-side additional condition should be studied. Instead of informing actual configuration, some kind of configuration ID and /or change timing of NW-side additional condition is necessary.**

Lenovo

1. Support procedures/signaling enabling UE/NW to associate the data/samples with the conditions/additional conditions under which the data/samples has been collected.
2. Support procedures/signaling enabling UE/NW for transmission of subset of samples among the set of measured/collected samples from the environment.
3. For transmission of ground-truth CSI samples, consider the performance of transmitting more samples, instead of fewer samples with higher resolution per sample (e.g., more samples with current parameter configurations for Rel-16 Type II, instead of less samples with a new parameter configuration for Rel-16 Type II), especially for cases that the overhead is more important, e.g., ground-truth data transfer for model monitoring or model update.

Samsung

**Proposal#9: For the cases of CSI compression with temporal aspects, consider the following for the network’s ground-truth CSI collection**

* **For cases that require multiple time-domain samples for inference, cases 2/3/4/5, high resolution codebook quantization including temporal aspects, e.g., Rel-18 eType II-like method with new parameters.**
* **For cases with CSI prediction, e.g., cases 3/4, high resolution codebook quantization for explicit channel matrices, e.g., codebook to report the left and right eigenvectors of a channel matrix**
* **Specification impact on measurement and reporting for ground-truth CSI**

**Proposal#10: For NW-first training with Option 4 (standardized data / dataset format + dataset exchange between NW-side and UE-side) study the necessity and specification impact of indication on the network-side additional condition.**

* **For two-sided models development, NW-part of two-sided model associated with a dataset can be considered as NW-side additional condition.**

ETRI

**Proposal 6: For dataset delivery for training collaboration type 3, for CSI compression sub-use case using two-sided model, consider the limited number of dataset samples to assess the feasibility of incorporating standardized signaling.**

MTK

1. For NW-side AI/ML model training, NW can rely on UL CSI samples collected from SRS sent by UEs. Use of DL CSI can be limited to finetuning purposes.

Qualcomm Incorporated

1. The triggering and / or configuration of UE side data collection should consider

* ***Configuration of associated ID for NW side additional information***
* ***Associated CSI-RS configuration***
* ***Standardized reporting of ground-ruth and its relevant capability is not needed***

CEWiT

**Proposal-11: For NW sided data collection, specify the CSI configuration, parameter combinations and periodicity for data collection.**

**Proposal-12: For UE sided data collection, existing procedures can be reused for data collection.**

**Proposal-13: For NW sided data collection, additional information e.g. SINR (in terms of CQI) on top of ground truth CSI.**

**Proposal-14: Consider channel parameter as a part of dataset-ID for data collection**

## Discussion

|  |  |
| --- | --- |
| Companies | views |
| Futurewei | Via High-res eT2 |
| Huawei | Confirm the necessity and feasibility of UE report ground-truth   * High-res eT2, NW configured rank, limit number of subbands   For type 3 training,   * study OTA signaling including dataset ID and dataset size, * study solution to relieve overhead (sent to multiple UE and gather at UE server) |
| Google | NW configured rank  Report singlular values, CQI/RI along with ground-truth  For UE side data collection, support configuration by NW and UE request |
| Tejas Networks | Prioritize precoder over raw channel |
| CMCC | Consider ground-truth reporting with existing eT2 or R18 doppler codebook, FFS potential enhancement for high-res data collection |
| ZTE | Via high-res eT2  Further report data quality, e.g., CQI / SINR  NW configures threshold for data quality |
| OPPO | NW side data collection study   * Dedicated RS or not, cell-specific or UE specific, perf vs. overhead   Consider condition information and additional information for data collection   * Condition can be CSI type / configuration * Additional condition can be scenario/cell ID, UE ID, LOS/NLOS flag |
| Fujistu | Via high-res eT2 |
| CATT | L1 or RRC signaling can be used, consider CSI framework if L1 signaling  Whether multiple CSI in same report for temporal case 2/3/4  Precoder is preferred over channel  NW configured rank |
| Panasonic | U-plan, RRC, MACCE sufficient for latency requirement of no-real time data collection  High-res eT2 can be studied  For NW side data collection, UE report associated ID for Rx filter assumption.  For UE side data collection, share NW side additional condition should be studied, via configuration ID and/or change timing of NW side additional condition. |
| Lenovo | Procedures / signaling enabling NW/UE to associate data samples with additional conditions /conditions  Consider UE/NW transmit subset of all samples  Consider transmit more data with eT2 than less data with high-res eT2 |
| Samsung | Consider high-res ground-truth for case 2/3/4/5, e.g., enhanced R18 doppler codebook  For cases w/ CSI prediction, consider ground-truth reporting of left/right singular vectors of channel  For inter-vendor collab option 4, study necessity and feasibility of indicating NW side additional condition |
| ETRI | Consider limited number of samples to assess the feasibility of incorporating standardized signaling |
| MTK | Consider data collection via SRS |
| QC | for UE side data collection,   * Consider configuration of associated ID for NW side additional information * Standardized Ground-truth reporting is not needed * Associated CSI-RS configuration |
| CEWiT | Specify CSI configuration, parameter combination, periodicity  Existing procedure can be used for UE side data collection  Report additional information, e.g., SINR along with ground-truth |

Data format, e.g., eT2, high-res eT2

Configuration and / or additional information

Signaling aspects

Association / configuration of additional condition

UE side data collection mechanisms

### NW-side data collection

Proposal 31a:

Confirm necessity and feasibility of ground-truth reporting for NW data collection for training. Consider following spec impacts

* Data format: codebook-based eT2 or R18 eT2.
  + FFS if enhancement is needed.
  + FFS number of samples in the report.
  + FFS whether channel or precoder is needed for temporal Cases 3 and 4
* Configuration of rank, number of subbands, CSI quality threshold
* Report CSI quality
* Report associated ID that captures UE side additional condition
* Configuration / reporting of temporal aspects for temporal case 2/3/4/5, e.g., association between input and output CSI
* Mechanism for data collection. Mechanisms outside L1-signlaing (e.g., RRC, U-plane, etc.) can be studied by other working groups.

|  |  |
| --- | --- |
| *Support / Can accept* |  |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Supportive of the direction.  We are a little bit hesitant to touch on the mechanism outside L1-signaling. It may be enough to focus on the high level:  “Mechanism for data collection, including signalling for the report. ~~Mechanisms outside L1-signlaing (e.g., RRC, U-plane, etc.) can be studied by other working groups.~~”  For “report of associated ID that captures UE side additional condition”, it is not fully understood how it works, we prefer either to delete it for now or we make it a high-level abstraction rather than directly goes to associated ID. |
| Xiaomi | We are general fine with the proposal.  For the fourth bullet, how to report UE side additional condition have not been discussed. It may be not need associated ID. We prefer to the associated ID replaced with **associated information** to make it be high level in current stage. |
| Huawei, HiSilicon | Several comments:  1) The eT2 like CB with new parameters should be part of the confirmation – legacy eT2 cannot provide precise enough label; if even the label cannot outperform eT2, the recovery CSI cannot outperform eT2 most likely. We may further study the methods to alleviate UE complexity, e.g., limit the subband number for per report, etc.  2) For CSI quality, fine with the direction; but we understand that SINR/CQI are already supported by legacy. Does it intend to introduce new quality types? Can be tentatively put to FFS and wait for more clarifications. Similarly, for associated ID provided by UE to NW, it can be put to FFS before further clarified.  3) For data collection signalling, it is not clear how UP signalling is workable as it may not be accessed by RAN.  **Suggested changes:**  Confirm necessity and feasibility of ground-truth reporting for NW data collection for training. Consider following spec impacts   * Data format: codebook-based eT2 with new parameters or R18 eT2 with new parameters.   + FFS if other enhancement is needed.   + FFS number of samples in the report.   + FFS whether channel or precoder is needed for temporal Cases 3 and 4   + FFS the enhanced values of parameters, and methods to alleviate UE complexity * Configuration of rank/layer, number of subbands,   + FFS CSI quality threshold * FFS Report CSI quality * FFS Report associated ID that captures UE side additional condition * Configuration / reporting of temporal aspects for temporal case 2/3/4/5, e.g., association between input and output CSI * Mechanism for data collection. Mechanisms outside L1-signlaing (e.g., RRC~~, U-plane, etc.~~) can be studied by other working groups. |
| Lenovo | Support with the following suggestion:   * Report associated ~~ID~~ information that captures other information related to the collection samples ~~UE side additional condition~~ |
| NEC | We suggest the following update:  Configuration of rank, configuration of layer, number of subbands, CSI quality threshold |
| Panasonic | We are fine with the proposal. |
| Futurewei | In general, we are ok with the direction with a few comments:   * + As what CSI quality means should be clarified first, we suggest adding “FFS” for this.   + UE side additional condition is still FFS as of RAN1#117, under the AI/ML model and data agenda item; thus, we suggest adding “FFS” to this bullet.   + In the last bullet, “Mechanisms outside L1-signlaing (e.g., RRC, U-plane, etc.) can be studied by other working groups.” can be moved to notes as it’s outside RAN1 and remove the “e.g.” part. |
| Samsung | OK in general. However, it is not clear whether indication of UE-side additional conditions via an associated ID is feasibility and necessary. The network is not expected to train AI/ML model for each corresponding particular implementation of UE-vendor. It is rather efficient, if the UE-side take care of its own additional condition after receiving training dataset or reference model in NW-first training. |
| LG | Generally fine. But listing the report associated ID without FFS is concerned as many companies mentioned |
| TCL | Agree. One way to enable the associated ID capture UE side additional condition is to encode the IDs based on the condition. |
| ZTE | OK with the direction. We also agree with HW that some issues can be added FFS since there is no full discussion on these issues, e.g., associated ID. |
| Ericsson | Support changes provided by Huawei. |
| Spreadtrum | For the fourth sub-bullet, associated ID is used to consistent NW additional condition. Thus, whether associated ID will capture UE side additional condition should be further studied. |

### UE-side data collection

Proposal 32a:

Confirm the necessity and feasibility of UE side data collection. Consider following spec impacts

* NW configuration or UE request for UE side data collection
* Configuration of temporal aspects for temporal case 2/3/4/5, e.g., association between input and output CSI
* Configuration of associated ID that captures NW side additional condition
* Note: data collection signaling / reporting, data format and associated information is up to UE-side implementation

|  |  |
| --- | --- |
| *Support / Can accept* |  |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Support |
| Xiaomi | Fine.  For the main bullet, we think UE side data collection is also used to training model. This should be clariid in the main bullet. |
| Huawei, HiSilicon | Not support – as we have been focusing on NW-side training and NW-first training, why do we still need UE to collect raw data? Regardless of option 3a/4/5a which may need dataset delivery, UE side can receive the collected/post trained dataset delivered by NW side. |
| Lenovo | Support |
| Panasonic | We are fine with the proposal. |
| Samsung | Ok in general. We propose to change data by dataset as this corresponds to NW-first training. We do not think the following is up to UE’s implementation   * ~~Note: data collection signaling / reporting, data format and associated information is up to UE-side implementation~~   We rather prefer to be modified as follows   * Note: Delivery of dataset from UE to a training entity is up to UE-side implementation |
| TCL | Agree. One way to enable the associated ID capture NW side additional condition is to encode the IDs based on the condition. |
| ZTE | We also have concerns about the proposal. For the 3rd bullet, we didn’t have any discussion on the configuration of associated ID in CSI agenda before, so we suggest deleting it. |
| Ericsson | Not support. This aspect has already been captured in study aspects on exchanging parameter/dataset/additional information from NW-side to UE-side for different inter-vendor training collaboration options with NW-first training. |

### Others

Please provide any other comments for this section.

|  |  |
| --- | --- |
| *Company* | *Comments* |
|  |  |
|  |  |

# Monitoring

## Summary of company proposals

From the submitted contributions, proposals related to monitoring are summarized below.

Futurewei

***Proposal 4: In AI/ML-based CSI compression using two-sided model, for NW-side monitoring based on the target CSI reported by the UE, consider adopting Rel-16 eType II CB based quantization with new parameters to achieve better monitoring performance.***

***Proposal 5: In AI/ML-based CSI compression using two-sided model, for UE-side monitoring, if the CSI reconstruction model or a reference CSI reconstruction model is available at UE, support at least the following:***

* ***Based on the output of the CSI reconstruction model at the UE (Case 2-1)***

***Proposal 6: In AI/ML-based CSI compression using two-sided model, for UE-side monitoring, further study the LCM complexity associated with using a proxy reconstruction model first before discussing whether to support the use of proxy reconstruction model at UE side for performance monitoring.***

Huawei

***Proposal 9: For NW-side monitoring, consider ground-truth CSI based monitoring in Rel-19.***

* ***eT2-like high-resolution codebook for reporting format with higher priority.***
* ***SGCS for the type of intermediate KPI with higher priority.***
* ***Further discuss the reporting mode, e.g., per sample reporting and reporting of a number of monitored samples.***

***Proposal 10: For UE-side monitoring, consider reconstructed CSI based monitoring (based on the output of the CSI reconstruction model indicated by the NW) in Rel-19.***

* ***eT2-like high-resolution codebook for indication format with higher priority.***
* ***SGCS for the type of intermediate KPI with higher priority.***
* ***Further discuss the indication mode, e.g., indication for a number of monitored samples.***

***Proposal 11: There is no strong motivation for specifying the UE side proxy model (Case 2-1/2-2) for monitoring.***

***Proposal 12: For UE-side monitoring, consider precoded RS (transmitted from NW based on the output of the CSI reconstruction model) based monitoring in Rel-19.***

* ***Further study the type of monitoring KPI calculated from the precoded RS.***

Spreadtrum, BUPT

***Proposal 3: For performance monitoring, support the following monitoring options.***

* ***NW side monitoring based on the ground-truth CSI reported by UE.***
* ***UE side monitoring based on the recovery CSI indicated by NW.***

Google

***Proposal 4: For CSI report based performance monitoring, the following spec impact should be considered:***

* ***New report quantity PMI only should be introduced, where UE reports the PMI based on a configured rank***
  + ***Support a further enhancement to report subband L1-SINR in addition to the PMI to facilitate the precoder and MCS selection for PDSCH***
* ***With regard to measurement accuracy in low SINR case, support the UE to report a state of CSI indicating the CSI is invalid for performance monitoring***
* ***With regard to joint ML based CSI compression and prediction, support to configure whether the UE should perform the CSI predication based on ML or non-ML and the CSI quantization based on ML or non-ML for separate performance monitoring for ML based CSI prediction and ML based CSI compression.***

***Proposal 5: For SRS based performance monitoring, the following spec impact should be considered:***

* ***Support to configure the SRS linked with a CSI-RS report configuration for ML based CSI, where the UE uses the same ports including antenna virtualization scheme to transmit the SRS and to receive the CSI-RS for ML based CSI***
* ***Support burst based SRS with frequency hopping to facilitate the performance monitoring for wideband channel for coverage-limited UE***

***Proposal 6: For UE-side monitoring, the following spec impact should be considered:***

* ***Introduce the hypothetical BLER as the metric for performance calculation***
* ***Configuration of precoded CSI-RS for hypothetical BLER calculation***

Tejas Networks

**Proposal 13: For NW side monitoring consider the target CSI reported by the UE via legacy eT2 codebook or eT2-like high-resolution codebook (Case 1) for better monitoring accuracy.**

**Proposal 14: For UE side monitoring consider the Case 2-1 for better monitoring accuracy**

* + **Based on the output of the CSI reconstruction model at the UE (Case 2-1)**
    - **Note: CSI reconstruction model at the UE-side can be the same as the actual CSI reconstruction model used at the NW-side, a reference model provided by NW, or a proxy model developed by the UE side.**

CMCC

***Proposal 10: For performance monitoring, the following two options could be prioritized:***

1. ***NW-side monitoring based on the ground-truth CSI report.***
2. ***UE-side monitoring based on the recovery CSI indication.***

ZTE

***Proposal 18:*** *Prioritize to study the specification impacts on at least the following case for model performance monitoring,*

* *NW-side monitoring based on the target CSI with realistic channel estimation associated to the CSI report, reported by the UE.*

***Proposal 19:*** *In CSI compression using two-sided model use case, deprioritize the study on UE-side monitoring in Rel-19 study phase.*

Ericsson

1. Conclude that it is necessary to specify UE reporting high resolution target CSI to enable NW-side intermediate KPIs based performance monitoring and performance degradation error cause detection for two-sided CSI-compression use case.
2. In CSI compression using two-sided model use case, capture in TR that ground-truth CSI report based on enhancements of the eType-II format with new parameters shall be defined to ensure high-accuracy model performance monitoring and error cause detection at the NW-side. Potential specification impact include:
   * Define the target-CSI format (e.g., Rel16 eType II CB with new parameters) for NW-side data collection (can reuse the ground truth defined for model training data collection)
   * Mechanisms (e.g., RRC-message based methods) to support UE reporting the target CSI together with the encoder output for NW-side data collection for performance monitoring.
   * Signaling and configuration for event triggered and periodical data collection at the NW-side.
3. For CSI compression using two-sided model use case, for any UE-sided performance monitoring method (if its feasibility and performance are justified), to enable the testability of the quality of the UE reported monitoring metrics, at least the following spec impact are identified:
   * The format of the monitoring metrics
   * Singaling and mechanisms for UE reporting monitoring metrics
   * RAN4 performance testing of the reported monitoring metrics
   * Data collection at the NW-side based on UE reporting monitoring data samples including target CSI, encoder output and monitoring metrics (to enable NW-side test the quality of the reported monitoring metric in the field).

Vivo

1. **For Option 3b (and potentially Option 3a/5a/4, if needed), the following methods can be considered to identify the cause of the performance degradation assuming NW-side monitoring**

* **Identifying data drift: with ground-truth CSI reported from UE, NW could check the SGCS of NW-side original model**
* **Identifying UE side cause:** 
  + **With ground-truth CSI reported from UE, NW compares expected CSI feedback and reported CSI feedback (FFS the necessity of taking quantized ground-truth as UE side input), and/or**
  + **NW compares Intermediate KPI of ongoing model and NW-side original model**
* **Identifying NW side cause: NW could check whether NW side decoder is correctly deployed by comparing the expected output and real output given the same decoder input**

1. **For UE-side monitoring, towards identifying the cause of performance degradation, we have the following considerations (assuming NW first training for option 3/4/5)**

* **Identifying UE side cause:** 
  + **For option 3a/5a-1 and 3b, UE could check the expected PMI of transferred CSI generation model and deployed CSI generation model**
  + **If original CSI generation model is not available at UE side (e.g., option 3a/5a-2 and option4), assistance information such as testing dataset from NW can help to identify UE side cause**
* **Identifying NW side cause: NW could check whether NW side decoder is correctly deployed by comparing the expected output and real output given the same decoder input**
* **If no UE side cause or NW side cause are identified, it can be determined that the performance degradation is due to data drift.**

1. **Both NW side and UE side monitoring can be supported for CSI compression with two-sided models**
   * + - **For NW-side monitoring, consider target CSI reporting by UE via legacy eT2 codebook or eT2-like high-resolution codebook ;**
       - **For UE-side monitoring, consider monitoring based on either direct estimation of SGCS (Case 2-1) or the output of reconstruction model at the UE (Case 2-2);**
         * **For provision of the reference model for UE-side monitoring purpose, it could directly follow the corresponding inter-vendor collaboration methods for Option 3, e.g., proxy model transfer or directly include monitoring related output as the model output.**
2. **Monitoring methods (both NW and UE side) can help to facilitate inter-vendor training collaborations, e.g., thorough identifying the cause of performance degradation to alleviate the concerns on the mismatch between transferred models and deployed models in option 3b.**

Xiaomi

***Proposal 7: Assume that only CSI generation model is deployed by using Option1/3/5 and the model is not retrained at UE side, the following mechanism could be considered to identify the cause of the performance degradation:***

* ***Step 1(applicable to Option 3/5 ): For the received model without retraining, the same target CSI is used as the input of CSI generation model both at UE side (e.g., OTT server) and gNB side, and CSI feedback generated by CSI generation model at UE side is sent to gNB.***
* ***Step 2: The same target CSI is used as the input of CSI generation model both at UE and gNB side, and CSI feedback generated by CSI generation model at UE side is sent to gNB.***

***Proposal 8: Assume that only CSI generation model is deployed by using Option 3/4/5 and the model is retrained at UE side, the following mechanism could be considered to identify the cause of the performance degradation:***

* ***Step 1 (applicable to Option 3/5 ): For the received model without retraining, the same target CSI is used as the input of CSI generation model both at UE side (e.g., OTT server) and gNB side, and CSI feedback generated by CSI generation model at UE side is sent to gNB.***
* ***Step 2: The target CSI applied for retraining AI/ML model is used as the input of retrained CSI generation model at UE side (e.g., OTT server) and CSI generation model at gNB side, and CSI feedback generated by CSI generation model at UE side is sent to gNB.***

***Step 3: The same target CSI is used as the input of CSI generation model both at UE and gNB side, and CSI feedback generated by CSI generation model at UE side is sent to gNB***

***Proposal 9: For different options of model performance monitoring, the following text proposal could be considered to be captured in TR 38.843.***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Monitoring option** | | **Overhead** | **Latency** | **Complexity** | **Monitoring accuracy** | **UE capability** | **Feasibility** |
| **NW-side monitoring** | **Based on the target CSI reported by the UE via legacy eT2 codebook or eT2-like high-resolution codebook** | **More** | **Depends on the number of reported target CSI** | **Higher** | **Reliable** | **Require to support eT2 codebook** | **Yes** |
| **SRS-based monitoring** | **Less** | **Depends on the number and interval of SRS transmission** | **Lower** | **Needs to evaluate** | **Basic feature** | **Needs to evaluate** |
| **UE-side monitoring** | **Option1: Based on the output of the CSI reconstruction model at the UE** | **Less** | **Depends on the number of output** | **Depends on the used model and increase LCM complexity** | **Depends on the used model** | **Require to support multiple models** | **Depends on the used model** |
| **Option2: Via direct estimation of intermediate KPI (e.g., SGCS) without reconstructing a target CSI** | **Less** | **Depends on the number of estimation of KPI** | **Depends on the estimation algorithm** | **Needs to evaluate** | **Require to support the estimation algorithm of KPI** | **Needs to evaluate** |
| **Option3: Via estimation of monitoring output other than intermediate KPI without reconstructing a target CSI** | **Less** | **Depends on the number of estimation of monitoring output** | **Depends on the estimation algorithm** | **Needs to evaluate** | **Require to support the estimation algorithm of monitoring output** | **Needs to evaluate** |
| **Option4: Based on precoded RS (e.g., CSI-RS, DMRS) transmitted from NW based on the output of the CSI reconstruction model** | **Less** | **Depends on the number of precoded RS and interval of adjacent precoded RS.** | **Lower** | **Needs to evaluate** | **Basic feature** | **Needs to evaluate** |
| **Option5: Based on the output of the CSI reconstruction model indicated by the NW via legacy eT2 codebook or eT2-like high-resolution codebook** | **More** | **Depends on the number of output of CSI reconstruction model** | **Higher** | **Reliable** | **Require to support eT2 codebook** | **Yes** |

Fujistu

***Proposal 13:***

* *For CSI compression using two-sided AI/ML models, the feasibility, reliability, and generalization capability of the UE-side AI/ML model performance monitoring using proxy model(s) should be evaluated and concluded before any further discussion on the related specification impacts.*

***Proposal 14:***

* *For the NW-side AI/ML model performance monitoring for CSI compression, RAN1 to prioritize the study of using the codebook-based quantization method to obtain the ground-truth CSI. Besides, adding new parameter values to legacy codebook for higher resolution ground-truth CSI should be studied.*

***Proposal 15:***

* *For CSI compression using two-sided AI/ML models, RAN1 to study the signaling and configuration for NW-side AI/ML model performance monitoring.*

***Proposal 16:***

* *For CSI compression using two-sided AI/ML models, regarding the NW-side AI/ML model performance monitoring using an existing CSI feedback scheme as a reference, RAN1 to study the potential specification impacts for the following three options:*
  + *Option-1: UE selects and reports PMI to the NW.*
  + *Option-2: UE computes and reports the intermediate KPI for the reference scheme, e.g., the SGCS of the recovered CSI from PMI and the ground-truth CSI.*
  + *Option-3: NW selects the PMI based on the ground-truth CSI reported by a UE.*

***Proposal 17:***

* *For CSI compression using two-sided AI/ML models, RAN1 to study the procedures and signaling needed for the follow-up actions after the AI/ML model performance monitoring, including falling back to legacy codebook-based CSI reporting from AI/ML-based methods.*

***Proposal 18:***

* *For the performance monitoring of CSI compression using two-sided AI/ML models, RAN1 to study the potential specification impacts on monitoring the performance of an inactive AI/ML model, taking at least the following cases into consideration:*
  + *Initial activation of an AI/ML model.*
  + *Re-activation of an AI/ML model.*

CATT

**Proposal 16: In CSI compression using two-sided model use case, for NW-side monitoring Case 1, further study the signalling and procedures for reporting target CSI, with the following two options considered:**

* **Option 1: The target CSI is reported separately from its associated CSI report;**
* **Option 2: The target CSI is reported together with its associated CSI report.**

**Proposal 17: In CSI compression using two-sided model use case, performance monitoring at UE-side based on reference model or proxy model can be deprioritized.**

**Proposal 18: In CSI compression using two-sided model use case, if eventual KPI is adopted as monitoring metric, how to exclude the impacts of other factors other than AI/ML model performance should be studied.**

**Proposal 19: In CSI compression using two-sided model use case, support UE-side monitoring based on precoded RS (e.g., CSI-RS, DMRS) transmitted from NW based on the output of the CSI reconstruction model.**

**Proposal 20: In CSI compression using two-sided model use case, for temporal domain aspects Case 3 and Case 4 with separate prediction and compression adopted, support monitoring the performance of the model for prediction and the performance of the model for compression separately.**

Panasonic

**Observation 24: Further study Direction 1 and Direction 3 with proxy model framework.**

* **Direction 1: Network-side monitoring based on the target CSI with realistic channel estimation associated to the CSI report, reported by the UE or obtained from the UE side.**
* **Direction 3: UE-side monitoring based on the output of the CSI reconstruction model at the UE side.**
  + **The CSI reconstruction part for performance monitoring at the UE is a proxy model, which is different from the actual CSI reconstruction part at the network.**

TCL

***Observation 4: The CSI report or indication for model monitoring in AI/ML based CSI compression introduces considerable overhead, RAN 1 should strive for an efficient signaling and overhead reduction mechanism for the CSI transmission.***

***Observation 5: The reliability of the overhead reduction scheme for CSI transferring for model monitoring in AI/ML based CSI compression should be guaranteed.***

Lenovo

1. In options 3a-1, 3b, 4-1, 5a-1, 5b, study mechanisms to ensure the applicability of the information received for the reference encoder model encoder model for the current input statistics.
2. Study mechanism to determine the main contributor(s) of the lower performance of the model at least for the issues related to a) the “deployed” encoder and/or decoder model, b) the communication link, or c) data-drift of the input data. We note that, in general, the “deployed” encoder/decoder model could have different performance that the “trained” (reference) encoder or decoder model.
3. Study mechanism for root-cause determination based on exchange of some test data-set between the NW and the UE.
4. Study mechanism for root-cause determination based on exchange of information regarding the NW-side trained encoder and/or decoder model.

Nvidia

**Proposal 6: RAN1 to study post-deployment performance monitoring mechanisms to detect performance degradation and non-compliance to guarantee satisfactory performance of AI/ML-based CSI compression in the field.**

InterDigital

**Proposal 1: Study further the following aspects for model monitoring in Rel-19:**

* **Details of reporting mechanism for the monitoring metrics with both time/event-trigger based**
* **Appropriate UE-side monitoring metric which reflects AI/ML model performance accurately**
* **UE-side monitoring based on precoded RS (CSI-RS, DM-RS)**
* **Reporting contents/structure of UE-side monitoring metric and its associated feedback overhead**
* **NW-side monitoring with lower signaling overhead**

**Proposal 2: Mechanisms for identifying the cause of performance degradation include:**

* **UE assistance information such as UE-side model monitoring metrics and UE-side out-of-distribution metrics**
* **Reporting the UE-assessed error cause (e.g., data drift, UE-side, or undetermined)**
* **Mitigation mechanisms, including fallback to legacy CSI reporting, model switching**

NEC

***Proposal 5: Support NW-side monitoring*** ***based on the target CSI reported by the UE via legacy eT2 codebook or eT2-like high-resolution codebook.***

***Proposal 6: For NW-side monitoring, the AI CSI and associated target CSI can be reported in the same reporting instance, or two separate reports.***

***Proposal 7: Support UE-side monitoring, the following can be considered:***

* ***Based on the output of the CSI reconstruction model at the UE. Where the CSI reconstruction model at the UE-side can be the same as the actual CSI reconstruction model used at the NW-side, a reference model provided by NW, or a proxy model developed by the UE side.***
* ***Via direct estimation of intermediate KPI (e.g., SGCS) without reconstructing a target CSI.***

Nokia

Proposal 7: Add the following methods to the list of options considered for network-side performance monitoring of CSI compression:

* **Direct estimation of an intermediate KPI based on encoded CSI messages**
* **Statistical monitoring based on pseudo-randomly subsampled elements of CSI information**

Proposal 8: In order to support root-cause determination of model failure, 3GPP should consider investigating feasibility of AI/ML-assisted root cause identification method.

Samsung

**Proposal#11: For performance monitoring, consider causes of performance loss that particularly affect the AI/ML based approach.**

* **Consider KPI for monitoring such as , where and are the SGCS for AI/ML-based CSI and baseline CSI, e.g., eType-II, respectively.**
* **To evaluate the monitoring options, apply the evaluation mechanism agreed in Rel-18 by replacing intermediate KPI by .**
* **FFS: baseline CSI for a payload size.**

**Proposal#12: For NW-side monitoring of two-sided models, when the input CSI is the (W2) domain, consider the same SD, FD and/or DD basis vectors for the CSI report for inference and monitoring purposes.**

**Proposal#13: To assess the accuracy of NW-side monitoring of two-sided models, when the input CSI is in the (W2) domain, KPIGenie is calculated with ground-truth CSI of Float32 representation of the W2 matrices.**

ETRI

**Observation 4: Regarding performance monitoring for CSI compression sub-use case using two-sided model, for Case 2-1:**

* **Computation complexity depends on the CSI reconstruction model on UE**
* **No communication complexity**
* **Latency depends on the inference time of the CSI reconstruction model on UE**
* **Accuracy depends on the difference between the reconstructed CSI and the actual output-CSI-NW. When UE uses the actual mode, the accuracy is the upper bound.**

**Proposal 7: For the performance monitoring of CSI compression sub-use case using two-sided model, conclude that Case 2-1 is feasible with considerations of complexity, latency, and accuracy.**

MTK

1. Discuss I/O-based monitoring, its relevant metrics, and benefits it offers compared to other monitoring methods.
2. Discuss NW-side AI/ML model monitoring using uplink CSI samples collected from SRS.
3. Discuss the NW-side monitoring with only considering the existing resolutions of CSI.

Apple

**Proposal 5: For case 3 of time-frequency-spatial domain CSI compression, CSI measurement in prediction window is the target CSI for NW side or UE side performance monitoring. The intermediate KPI or eventual KPI includes both compression and prediction performance.**

**Observation 6: NW side performance monitoring has issue of high feedback overhead and additional UE complexity and power consumption for extended parameter sets.**

**Observation 7: For UE side performance monitoring using proxy model, since inter-vendor collaboration option 3a-2, 3a-3, option 5a-2, 5a-3, and option 4-2 and option 4-3 provide information related to CSI reconstruction model, the proxy model can be trained based on this information. However, NW may need to perform performance monitoring on the proxy model.**

**Observation 8: For UE side performance, NW implicitly transmit output CSI using precoded CSI-RS to the UE provide a simple and low overhead solution.**

**Proposal 7: For CSI compression using two-sided model, for UE side performance, further study RLF/BFD like mechanism for UE initiated report.**

NTT Docomo

Table 2 Pros and Cons of Performance Monitoring Approaches

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Monitoring Scheme | | Overhead | Latency | Complexity | Accuracy | UE Capability |
| NW-side | eT2-like CSI based | UL overhead (CSI report) | Short | High: UE to support eT2 CB quantization. | Accurate | New cap. for eT2-like CB quantization |
| SRS based | UL overhead (SRS transmission) | Short | Minor w/ legacy channel estimation.  High w/ AI/ML- based UL-DL channel prediction. | Inaccurate w/ non-AI/ML channel estimation.  FFS AI/ML FDD channel estimation. | Existing cap. for SRS transmissions.  New cap. For SRS-based UL-DL channel prediction. |
| UE-side | Output CSI reconstruction | Minor | Short | Highest w/ CSI reconstruction model. | Depend on the model complexity and its generalization performance. | New cap. for estimation model |
| Direct KPI estimation | Minor | Short | High: UE to support KPI estimation model and corresponding LCM. | Need FFS (can be better than Output CSI reconstruction). | New cap. for estimation model |
| Direct monitoring output estimation | Minor | Short | High: UE to support monitoring output model and corresponding LCM. | Need FFS (can be better than Output CSI reconstruction). | New cap. for estimation model |
| Precoded RS | DL overhead (CSI-RS transmissions) | Long (round-trip) | Medium: UE additional RS estimation. | Accurate | Existing cap. for more channel estimation |
| eT2-like CB indicated from NW | DL overhead (CSI indication) | Long (round-trip) | High: UE to support eT2 CB reconstruction. | Accurate | New cap. for eT2-like CB reconstruction |

**Observation 3**

* **For the NW-side monitoring, the eT2-like CSI-based approach has better accuracy than the SRS-based one.**
* **For the UE-side monitoring, the direct KPI estimation can outperform the other regarding the overhead, latency, or complexity if its accuracy can be justified.**

**Proposal 3**

* **For NW-side monitoring, support the scheme based on the eT2-like CSI report from the UE.**
* **For UE-side monitoring, further study the reliability of direct KPI estimation for the down-selection among direct KPI estimation, precoded RS, and eT2-like indication.**

Qualcomm Incorporated

1. Consider two-phase mechanism for performance monitoring and identification of root cause

* Phase 1: identifying whether there is performance degradation due to AI/ML models
* Phase 2: identifying whether the performance degradation is due to UE side CSI generation part, NW side CSI reconstruction part, or data drift, etc.

1. Recommend following model performance monitoring mechanisms for normative work, and study potential specification impacts of triggering / configuration / reporting of the ground-truth

* NW side monitoring mechanism with ground-truth reporting
  + Phase 1 monitoring of performance degradation:
    - NW uses the reported ground-truth to assess the performance of actual CSI generation part and CSI reconstruction model
  + Phase 2 monitoring of root cause identification:
    - NW assess whether performance degradation is due to UE side issue, using the reported ground-truth, based on a reference CSI generation model at NW side
* UE side monitoring mechanism
  + Phase 1 monitoring of performance degradation:
    - UE generates SGCS using a SGCS estimator associated to its CSI generation part and report and reports the SGCS estimate
  + Phase 2 monitoring of root cause identification:
    - If the SGCS is bad, UE may further report ground-truth. NW assess whether performance degradation is due to UE side issue, using the reported ground-truth, based on a reference CSI generation model at NW side.
    - Note: the ground-truth reporting is far less frequent than SGCS report in phase 1.

CEWiT

**Proposal-15: For AI/ML based CSI compression, RAN1 to study the signaling and configuration for NW-side AI/ML model performance monitoring.**

**Proposal-16: For AI/ML based CSI compression, consider model monitoring via UE side proxy model with less priority.**

**Proposal-17: For AI/ML based CSI compression, consider model monitoring via estimation of intermediate KPI over proxy model-based monitoring.**

**Proposal-18: Study the feasibility of transmitting the strongest basis and its corresponding value instead of transmitting the entire input in case of model monitoring.**

**Proposal-19: For AI/ML based CSI compression, study signalling and follow-up procedure for the outcome of AI/ML model monitoring.**

## Discussion

|  |  |
| --- | --- |
| companies | views |
| Futurewei | NW side monitoring w/ target CSI reporting via eT2 or high-res eT2  UE side monitoring,   * Consider case 2-1 if NW decoder is available at UE side * ref reconstruction model or proxy need to study its LCM |
| Huawei | NW side monitoring,   * ground-truth reporting via high-res eT2, SGCS metric, indication of multiple samples   UE side monitoring,   * Consider reconstructed CSI indicated from NW via high-res eT2 * Consider precoded RS based solution, FFS monitoring KPI * No strong motivation for proxy model |
| Spreadtrum, BUPT | NW side monitoring w/ ground-truth reporting  UE side monitoring w/ reconstructed CSI indication |
| Google | NW side monitoring   * PMI-only report w/ a given rank, indication of CSI quality * Separate monitoring for predication and compression * SRS based monitoring (linkage between SRS and CSI, enh of burst SRS hopping)   UE side monitoring   * Precoded CSI-RS w/ hypo BLER metric |
| Tejas Networks | NW side monitoring w/ ground-truth reporting via eT2 or high-res eT2  UE side monitoring based on reference decoder |
| CMCC | NW side monitoring w/ ground-truth reporting  UE side monitoring w/ reconstructed CSI indication |
| ZTE | NW side monitoring w/ ground-truth reporting  Deprioritize UE side monitoring |
| Ericsson | Conclusion high-res eT2 ground-truth reporting is needed for performance monitoring and root-cause detection   * Define format, mechanisms, signaling / configuration for event trigger   For any UE side monitoring,   * Monitoring metric format, signaling / mechanism, RAN4 testing, data collection enable NW side test the quality of the monitoring metric |
| vivo | Explain both NW and UE side monitoring can identify the root cause of performance degradation  Both NW and UE side monitoring can be considered. NW side with ground-truth reporting, UE side with ref decoder or SGCS estimator |
| Xiaomi | Explain root cause identification can be achieved via inferencing using same input CSI for UE side actual encoder and NW encoder.  Propose comparison table for monitoring options |
| Fujistu | Feasibility, generalization (various scenario, different NW vendor’s decoder), reliability of UE side monitoring using proxy model should be evaluated before study spec impacts  NW side monitoring   * consider ground-truth reporting using eT2 or high-res eT2, * study signaling / configuration, * reporting content (UE reports PMI, UE reports PMI+SGCS, NW selects PMI based on ground-truth reported by UE)   Study follow-up actions, study monitoring of inactive model |
| CATT | NW side monitoring, study whether ground-truth is reported in same or separate CSI report as the ML CSI  Deprioritize UE side monitoring using ref decoder or SGCS estimator  Support UE side monitoring based on precoded RS  Support separate monitoring for compression and prediction |
| Panasonic | NW side monitoring w/ ground-truth reporting,  UE side monitoring w/ reference decoder |
| TCL | Overhead issue of reporting CSI, RAN1 strive to study overhead reductionscheme for model monitoring. |
| Lenovo | Study mechanism for root-cause detection considering exchange of some dataset  Study mechanism for root-cause detection considering exchange of NW side encoder and/or decoder |
| InterDigital | Study appropriate UE side monitoring metric and its reporting mechanism, UE side monitoring based precoded RS  Study NW side monitoring with lower signaling overhead  Root cause identification at UE side:   * Assistance information and UE side OOD metric, reporting the UE assessment, mitigation mechanisms including fallback |
| NEC | NW side monitoring w/ ground-truth reporting via eT2 or high-res eT2, whether it is reported in the same or separate report as the ML CSI.  UE side monitoring, consider ref decoder or SGCS estimation. |
| Nokia | NW side monitoring w/ direct estimation of KPI or subsampled target CSI reporting  For root cause detection, further consider the feasibility of AI/ML assisted method |
| Samsung | Propose new metric for monitoring derived from SGCS gap  NW side monitoring consider same SD-FD bases if ML CSI is in beam-delay domain |
| ETRI | Conclude UE side monitoring using ref decoder is feasible. |
| MTK | Propose to discuss input/output based monitoring  NW side monitoring using SRS, or existing codebook for ground-truth reporting |
| Apple | For temporal case 3, intermediate KPI includes KPI for both prediction and comporession.  Further consider RLF/BFD like mechanism for UE initiated report   * Ref decoder based UE side monitoring may need NW further monitoring * Precoded RS based UE side monitoring provides simple and low overhead solution |
| DCM | NW side monitoring based on eT2 like ground-truth reporting  SGCS estimator for UE side monitoring needs further study of its reliability |
| QC | Propose two-phase mechanism for performance monitoring and root cause identification.  Recommend NW side monitoring w/ ground-truth reporting for monitoring and root cause identification  Recommend UE side monitoring for performance monitoring + occasional ground-truth reporting for root cause identification |
| CEWiT | Study NW monitoring signaling and configuration  Study UE monitoring using proxy model with less priority  Consider UE side monitoring using intermediate KPI estimation  Study reporting strong basis rather than entire CSI for ground-truth reporting  Study follow up actions |

|  |  |  |
| --- | --- | --- |
| Topics | Supporters w/ potential spec impacts | Concerns or objections |
| NW side monitoring   * based on target CSI reporting **(18, 5 w/ conditions)** | Futurewei, Huawei, Google, Spreadtrum, Tejas Network, ZTE, Ericsson, vivo, Xiaomi, Fujistu, Panasonic, TCL (overhead reduction), InterDigital (lower overhead), NEC, Nokia (subsampled target CSI), Samsung, MTK (just eT2), DCM, QC (for root cause identification occasionally)   * eT2 or high-res eT2 (7) * define SGCS metrics * indication of multiple samples * PMI-only report for a given rank * Needed for root-cause identification * Mechanism * Signaling / configuration for event trigger * Reporting content * Lowering overhead | Concern:  Overhead, latency, no need of eT2 enhancment |
| NW side monitoring   * based on SRS **(3)** | Google, vivo, MTK   * Burst SRS hopping to cover wideband |  |
| UE side monitoring   * based on NW indication of reconstructed CSI **(4)** | Huawei, Spreadtrum, CMCC, vivo   * eT2 or high-res eT2 * monitoring metrics * indication of multiple samples * Signaling / mechanism * RAN4 testing * Data collection that enables NW testing | ZTE (deprioritize) |
| UE side monitoring   * based on precoded RS **(6)** | Huawei, vivo, CATT, InterDigital, Apple, QC   * FFS monitoring KPI (SGCS, hypo BLER) * Signaling / mechanism (e.g., RLF/BFD like) * RAN4 testing * Data collection that enables NW testing | ZTE (deprioritize)  Additional RS overhead, latency, intermediate KPI calculation is not available |
| UE side monitoring   * based on NW decoder or reference decoder **(8)** | Futurewei, Tejas Network, vivo, Panasonic, Lenovo, NEC, ETRI, QC   * Signaling / mechanism (e.g., RLF/BFD like) * RAN4 testing * Data collection that enables NW testing * Monitoring metric | ZTE, CATT, CEWiT (deprioritize)  Huawei (no strong motivation, additional LCM)  Fujistu (feasibility, generalization issue) |
| UE side monitoring   * based on SGCS estimator (direct KPI estimation) **(4)** | Vivo, NEC, QC, CEWiT   * Signaling / mechanism * RAN4 testing * Data collection that enables NW testing * Monitoring metric | ZTE, CATT  Huawei (no strong motivation, additional LCM)  Fujistu (feasibility, generalization issue)  DCM (reliability) |
| Other aspects | Follow up actions, e.g., fallback, switching  Separate monitoring for prediction and compression  Root cause detection via dataset sharing or encoder/decoder sharing  Root cause detection via OOD metric  AI/ML based root cause identification |  |

### Way forward for monitoring

FL observation

* There is majority (14) supportive of NW side monitoring with gourd-truth CSI reporting via eT2 or high-resolution eT2 codebook
  + 5 companies raise concern on incurred latency / overhead, and feasibility of high-resolution eT2.
* Also good number of companies favour UE side monitoring via precoded RS (6) or reference (or actual) decoder (8)
  + For reference decoder, there may be additional monitoring / LCM raised by some companies
* 4 companies support NW indication of reconstructed CSI
* 4 companies support using proxy or SGCS estimator that measure SGCS directly.
  + Main concern is reliability and generalization of such AI/ML model even though 1 company provides evaluation results showing good generalization ability across random vectors and Dense Urban scenario.
  + Another concern is additional monitoring / LCM
* 1 company mention that UE side monitoring methods needs to consider RAN4 testability and NW side data collection to enable NW side testing.

From FL perspective

NW side monitoring is feasible if overhead / latency issue can be addressed. One possible way is to combine with UE side monitoring to decrease the frequency of ground-truth reporting. Regarding whether high-resolution eT2 is needed, further evaluation is encouraged considering the monitoring accuracy per N samples instead of per single sample, i.e., |avg(SGCS\_groundtruth)-avg(SGCS\_actual)|. This is because in practical monitoring, the decision will be made by averaging the SGCS over N samples.

UE side monitoring using precoded RS is feasible if the monitoring metric issue can be addressed. There are multiple factors that may impact the performance of precoded RS vs. original ground-truth CSI, e.g., channel variation, channel estimation, etc.

UE side monitoring using proxy model or SGCS / monitoring output estimator, is feasible if the proxy model is the actual decoder used by NW. Otherwise, the main concern lies in additional LCM burden for the proxy / estimator model. To address this issue, RAN4 testing and data collection by NW to perform NW side testing of the proxy / estimator can be considered. Regarding the reliability / generalization performance, one company presents results showing good potential. FL encourages other companies to provide results so as to make conclusion.

Hence, FL proposes the below.

Proposal 41a:

For model performance monitoring, conclude that

* NW side monitoring with ground-truth CSI reporting from UE is feasible, under the condition that
  + Method of handling incurred overhead/ latency is identified (e.g., by combining with UE side monitoring methods, etc)
* Further study whether enhancement of eT2 is needed.
* UE side monitoring with the listed options is feasible under the condition
  + Proper monitoring KPI is identified, e.g., SGCS for methods based on CSI reconstruction model and SGCS estimator, other metrics for methods based on precoded RS or monitoring output
  + RAN4 testability is ensured and data collection procedure that enables NW to test the monitoring metric / output is enabled (e.g., ground-truth reporting from UE side).

Evaluate the accuracy, overhead, latency, complexity, and generalization ability of the NW-side and the UE-side monitoring approaches and combination of NW-side/UE-side approaches.

* Performance evaluation should consider calculating the monitoring KPI based on N samples. That is, KPI*Diff* = *f* ( KPI*Actual* , KPI*Genie* ), where KPI*Actual* is obtained based on N samples. In case of SGCS as KPI, KPI*Actual* can be derived by averaging the SGCS estimate over N samples. Likewise, KPI*Genie* is derived based on the same N samples.

|  |  |
| --- | --- |
| *Support / Can accept* |  |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Support |
| Xiaomi | We are fine with NW side monitoring in the proposal. For UE side monitoring, we have also concern on it feasible based on SGCS estimator. In addition to, it also incurs the more complexity of LCM as raised by other companies. |
| Huawei, HiSilicon | 1) For NW side monitoring, the key point of the feasibility is whether UE (considering its capability/complexity) can report ground-truth CSI is enhanced eT2 with new parameters. From performance perspective, it is clear that higher resolution benefits to monitoring accuracy. Let’s consider an extreme example, where we adopt legacy eT2 as the label fed back by UE, and assuming a benchmark of legacy eT2 as the non-AI solution, then from monitoring results at NW, benchmark performance is always 100%, and there is never a need to enable the AI solution.  2) For UE side monitoring, the key point of the feasibility includes: i) whether the complexity of managing proxy model (Case 2-1/2-2) is affordable by NW, or ii) whether the output of the proxy model, if not managed by NW, is trustable/testible.  3) For the new EVM, the monitoring methods include two aspects: one is the averaged monitoring, which monitors over a long time on its long term performance; the other is per sample based monitoring, which monitors the reliability of the model. From our perspective, both are valid, so it is suggested to keep both 1 sample and N samples in the EVM.  **Suggested changes:**  For model performance monitoring, conclude that   * The feasibility of NW side monitoring with ground-truth CSI reporting from UE depends on whether the ground-truth CSI is enhanced eT2 with new parameters ~~is feasible,~~ and under the condition that   + Method of handling incurred overhead/ latency is identified (e.g., by combining with UE side monitoring methods, etc) * ~~Further study whether enhancement of eT2 is needed.~~ * The feasibility of UE side monitoring with the listed options depends on ~~is feasible~~ ~~under~~ the following condition   + Proper monitoring KPI is identified, e.g., SGCS for methods based on CSI reconstruction model and SGCS estimator, other metrics for methods based on precoded RS or monitoring output   + Whether the NW management complexity to the proxy model (if needed) is affordable.   + RAN4 testability is ensured and data collection procedure that enables NW to test the monitoring metric / output is enabled (e.g., ground-truth reporting from UE side).   Evaluate the accuracy, overhead, latency, complexity, and generalization ability of the NW-side and the UE-side monitoring approaches and combination of NW-side/UE-side approaches.   * Performance evaluation should consider calculating the monitoring KPI based on both 1 sample and N samples.   + Monitoring KPI based on 1 sample is calculated same as R18 EVM.   + For N samples ~~That is~~, KPI*Diff* = *f* ( KPI*Actual* , KPI*Genie* ), where KPI*Actual* is obtained based on N samples. In case of SGCS as KPI, KPI*Actual* can be derived by averaging the SGCS estimate over N samples. Likewise, KPI*Genie* is derived based on the same N samples. |
| Lenovo | Support |
| NEC | Support |
| Panasonic | Support |
| Futurewei | We are ok with edits from Huawei, HiSilicon. |
| ETRI | Support the proposal. |
| Intel | At least the following aspects are not considered for the Rel-18 evaluation methodology: (1) robustness against channel variations in time, (2) associated overhead. So, we propose to remove the last sub-bullet. |
| LG | Fine |
| TCL | Agree |
| ZTE | Agree with HW’s modification.  In addition, for the 2nd part, we suggest adding a bullet **“Performance evaluation should consider calculating the monitoring KPI for rank>1**” since we only evaluate the monitoring performance based on rank=1. If rank>1, we didn’t have further discussion, but in practical, Rank2 is evaluated most for SGCS and UPT. So, the monitoring KPI for rank>1 is important to consider in further evaluation.   * Performance evaluation should consider calculating the monitoring KPI based on both 1 sample and N samples.   + Monitoring KPI based on 1 sample is calculated same as R18 EVM.   + For N samples ~~That is~~, KPI*Diff* = *f* ( KPI*Actual* , KPI*Genie* ), where KPI*Actual* is obtained based on N samples. In case of SGCS as KPI, KPI*Actual* can be derived by averaging the SGCS estimate over N samples. Likewise, KPI*Genie* is derived based on the same N samples.   + Monitoring KPI for rank>1 |
| Ericsson | Support modifications from Huawei and ZTE. |

### Root cause identification

FL observation

* 7 companies (Ericsson, vivo, QC, Xiaomi, Lenovo, InterDigital, Nokia) discuss and analyze methods of root cause identification
  + One company propose to conclude that NW side monitoring with ground-truth reporting can achieve root cause identification
  + 5 companies also discuss method of gournd-truth reporting, but raise overhead issue and performance accuracy.
  + One company discuss UE side monitoring approach and observe it is feasible to identify root cause
  + One company discuss possibility of identifying root cause based on UE feedback latent via AI/ML approach.
  + One company discuss potential spec impacts, e.g., configuration, mechanism, metric, etc

Proposal 42a:

Conclude that root cause identification can be achieved by ground-truth CSI reporting

* e.g., with the understanding that NW side runs inference using the NW side reference CSI generation part and compare with the inference using UE side CSI generation part.
* Note that this conclusion does not preclude other methods.

|  |  |
| --- | --- |
| *Support / Can accept* |  |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Support |
| Xiaomi | Support |
| Huawei, HiSilicon | Support. |
| Lenovo | We support this proposal with the following modifications:  Conclude that root cause identification can be achieved at least by ground-truth CSI reporting |
| NEC | Same view with Lenovo. |
| Panasonic | Support Lenovo’s update proposal. |
| Futurewei | We are ok with edits from Lenovo. |
| LG | Fine |
| TCL | Agree |
| ZTE | Support |
| Ericsson | It is unclear how root cause identification (e.g., NW side, UE side, data drift) can be done based on the simple example text in the proposal. More clarification is needed. The example shall also mention which inter-vendor training collaboration option(s) is assumed, which is important for making conclusions on the feasibility of different training options. |
| Spreadtrum | Support |

### Others

Please provide any other comments for this section.

|  |  |
| --- | --- |
| *Company* | *Comments* |
|  |  |
|  |  |

# Inference aspects (pairing / CQI / quantization)

## Summary of company proposals

From the submitted contributions, proposals related to inference aspects (pairing, CQI, quantization, etc.) are summarized below.

Futurewei

***Proposal 7: In AI/ML-based CSI compression using two-sided model, for CQI determination,*** ***if the CSI reconstruction model or a reference CSI reconstruction model is available at UE, adopt Option 2a to determine CQI at UE.***

***Proposal 8: In AI/ML-based CSI compression using two-sided model, the discussion of model pairing options and the associated procedures can be deferred till some consensus is reached among companies on other related and already on-going discussions, e.g., model identification options/procedures.***

Huawei

***Proposal 13: For quantization methods of the CSI report, further study potential specification impact on quantization alignment using standardized quantization scheme.***

* ***For vector quantization,***
  + ***Configuration/reporting/updating of the quantization dictionary.***
  + ***Segmentation of the CSI generation model output to map with short VQ vector.***
* ***For scalar quantization,***
  + ***The configuration of the quantization granularity/range.***

***Proposal 14: For the study of CQI determination in inference, consider Option 1 (CQI is NOT calculated based on the output of CSI reconstruction part from the realistic channel estimation) as a starting point.***

***Proposal 15: For CSI report in inference, on top of the legacy CSI reporting principles, the following AI/ML specific aspects may be additionally studied:***

* ***The CSI priority rules, e.g., priority rules by considering the AI/ML specific reporting type, priority rules within the bit sequence of per AI/ML specific inference CSI report.***
* ***The CSI processing unit (CPU), e.g., the required CPU value may consider difference of UE part model complexity.***
* ***The CSI mapping, e.g., factors representing the part 2 size in CSI part 1, mapping of the CSI generation part output in CSI part 2, etc.***

***Proposal 16: For Rank>1 options in inference, further study Option 3-1 (layer common and rank common)*** ***Option 3-2 (layer common and rank specific) and Option 2-1 (layer specific and rank common) with higher priority.***

* ***For Option 3-1, it is straightforward that the LCM of activation/deactivation/monitoring is applied to this single model.***
* ***For Option 3-2 and Option 2-1, need to further discuss whether to consider the multiple models as separate models or a single model from LCM perspective.***

Spreadtrum, BUPT

***Proposal 1: For the study of CQI determination in inference, support Option 1b (CQI is calculated based on target CSI with realistic channel measurement and potential adjustment).***

***Proposal 2: Consider to report historical CSI information via NW-triggered signaling when UCI missing or UCI dropping.***

Google

***Proposal 1: Support the following types of CSI report for CSI compression:***

* ***Type 1 (Compression of channel): UE reports subband L1-SINR and compressed channel***
* ***Type 2 (Compression of channel eigenvector): UE reports compressed channel eigenvector for a configured rank***
* ***Type 3 (Compression of W2): UE reports W1 and compressed W2 for a configured rank***

***Proposal 2: The priority for non-ML based CSI report should be higher than the priority of ML based CSI report.***

***Proposal 3: Support the CPU occupancy rule for ML based CSI based on two types processing unit***

* ***Type1 CPU: a measurement processing unit (MPU) used for channel estimation and pre-processing***
* ***Type2 CPU: an inference processing unit (IPU) used for inference for ML based CSI***

***Proposal 12: Support hybrid AI/ML based and non-AI/ML based CSI measurement and report***

* ***UE reports the CSI based on AI/ML if it reports a small RI and the UE can report the CSI based on Type1 codebook if it reports a large RI***

ZTE

***Proposal 17:*** *For CQI determination, at least prioritize the specification impact discussions on Option 1a, Option 1b.*

Xiaomi

***Proposal 10: The legacy priority rule can be reused to define the priority the AI/ML based CSI reporting, and a priority value with new parameter value or introducing new parameter is used to indicate the priority of AI/ML based CSI reporting.***

***Proposal 11: The compressed CSI part 2 should be divided into 1<N groups for CSI omission. How to divide compressed CSI part 2 into N groups needs to further study.***

***Proposal 12: If multiple predicted CSI of the multiple future instances are reported in one CSI reporting, how to pack the multiple CSI in the CSI reporting needs to study.***

***Proposal 13: If there is no output of historic CSI at the previous instance, how to design the current input of historic CSI for two-sided AI/ML model needs to study.***

Fujistu

***Proposal 5:***

* *For CSI compression using two-sided AI/ML models, support both the following alternatives of precoding matrix for output-CSI-UE and input-CSI-NW:*
  + *Alt 1: The precoding matrix in spatial-frequency domain*
  + *Alt 2: The precoding matrix represented using angular-delay domain projection.*

***Proposal 6:***

* *For CSI compression using two-sided AI/ML models, support the following approaches for AI/ML model alignment:*
  + *UE initiated: UE reports the pairing information for NW confirmation.*
  + *NW initiated: NW indicates the pairing information supported for UE confirmation.*
* *Pairing information could be in the form of model ID.*

***Proposal 7:***

* *For CSI compression using two-sided AI/ML models, RAN1 to further study using local model IDs in AI/ML model operations and CSI configuration/reporting after model alignment between UE and NW, which reduces the overhead compared to global model IDs.*

***Proposal 8:***

* *For CSI compression using two-sided AI/ML models, global model ID is sufficient for model alignment, and there is no need to introduce pairing IDs.*

***Proposal 9:***

* *For CSI compression using two-sided AI/ML models, RAN1 to further study the configurations and CSI reporting formats required for various AI/ML model settings. To reduce the normative workload, the following could be down selected:*
  + *AI/ML-model-setting-specific CSI configurations and CSI reporting formats.*
  + *A configuration and CSI reporting format adapting to various possibilities, including at least*
    - *layer specific and rank common.*
    - *layer specific and rank specific.*
    - *layer common and rank common.*
    - *layer common and rank specific.*

***Proposal 10:***

* *For CSI compression using two-sided AI/ML models, deprioritize Option 2 proposed in RAN1 #112 for CQI determination.*
  + *Option 2: CQI is calculated based on the output of CSI reconstruction part from the realistic channel estimation.*

***Proposal 11:***

* *In the CSI reports generated by AI/ML, include the information for the order of the spatial layers of the reported precoding matrix related information, if the reported rank is larger than 1.*
* *Furthermore, the information in the first bullet should be in the Part I CSI.*

***Proposal 12:***

* *In the CSI reports generated by AI/ML, the criteria of the Part II CSI priority can be set by the order of the spatial layers of the precoding matrix related information , if the reported rank is larger than 1.*
* *Furthermore, the order of the spatial layers in the first bullet can be*
  + *Alt 1: the same as the order of the reported layers of the precoding matrix related information.*
  + *Alt 2: the descending order of the singular value corresponding to the precoding vectors.*

CATT

Proposal 21: In CSI compression using two-sided model use case, quantization alignment between UE-side and NW-side in a 3GPP non-transparent manner is supported.

Proposal 22: In CSI compression using two-sided model use case, legacy CSI reporting principles are reused as much as possible.

Proposal 23: In CSI compression using two-sided model use case, if CQI in CSI report is configured, for CQI determination in CSI report, one of the sub options of Option 1 is adopted:

* **Option 1: CQI is NOT calculated based on the output of CSI reconstruction part from the realistic channel estimation, including**
  + **Option 1a: CQI is calculated based on target CSI with realistic channel measurement**
  + **Option 1b: CQI is calculated based on target CSI with realistic channel measurement and potential adjustment**
  + **Option 1c: CQI is calculated based on legacy codebook.**

Proposal 24: For CQI reporting in CSI compression using two-sided model use case, the same quantization scheme as that in Rel-17 for codebook based CSI feedback is considered.

Panasonic

**Observation 13: For CQI determination in CSI report, further study following options.**

* **Option 1: CQI is NOT calculated based on the output of CSI reconstruction part from the realistic channel estimation, including**
  + **Option 1a: CQI is calculated based on target CSI with realistic channel measurement**
  + **Option 1b: CQI is calculated based on target CSI with realistic channel measurement and potential adjustment**
* **Option 2: CQI is calculated based on the output of CSI reconstruction part from the realistic channel estimation, including**
  + **Option 2a: CQI is calculated based on CSI reconstruction output, if CSI reconstruction model is available at the UE and UE can perform construction model inference with potential adjustment**
    - **The CSI reconstruction part for CQI determination at the UE is a proxy model, which is different from the actual CSI reconstruction part at the network.**

TCL

***Proposal 1: In the case of CSI compression using a two-sided model, the design of an AI/ML-specific CSI-RS resource and CSI reporting configuration that may be compatible with the traditional CSI reporting scenario should be considered in the following aspects:***

* ***AI/ML-specific CSI-RS resource configuration for CSI compression***
* ***AI/ML-specific fields in CSI-ReportConfig IE***
* ***Dedicated report quantities and report configurations for AI***

***Proposal 2: The definition of AI/ML-specific priority for CSI reporting in relation to CSI compression should be considered in comparison with the traditional CSI priority rules.***

***Proposal 3: When the UE supports both AI/ML and non-AI/ML CSI reporting, it is necessary to redefine the priority rule considering different types of CSI reporting.***

***Proposal 4: The study of how to describe the capabilities of a UE to implement AI/ML models for inference on CSI compression and calculations should be undertaken.***

LGE

**Proposal #1: Regarding temporal/spatial/frequency (TSF)-domain CSI compression, study methods/mechanisms to manage the similarity/synchronization of accumulated past CSI at UE-side and/or NW-side.**

**Proposal #2: Regarding TSF-domain CSI compression, discuss the format of historical CSI information and how to report it at least for performance monitoring perspective.**

**Proposal #3: Regarding TSF-domain CSI compression Case 3 and 4, consider performance monitoring method on joint CSI compression and prediction by adapting the operation on the AI/ML model between CSI compression and prediction.**

**Proposal #4: Regarding non-ideal UCI feedback on TSF-domain CSI compression,**

* **Consider two-step performance monitoring to check that the performance degradation of the AI/ML model is originated from whether the historical CSI has a problem or the AI/ML model is not suitable for the deployed environment**
* **Also consider to report past CSI information via NW-triggered signaling when UCI missing or UCI dropping.**

**Proposal #5: Consider the method on the rank adaptation based on the validity check of layer(s) for a given RI.**

**Proposal #6: For CQI determination in CSI compression using two-sided model, consider to prioritize Option 1. If Option 2 is supported, further consider**

* **Option 2a: Utilizing AI/ML model complexity reduction method to reduce the signaling overhead to deliver the CSI reconstruction part at NW-side.**

Lenovo

1. Support procedures/signalling enabling CSI-compression models having both Scaler and vector Quantizers for generation of the CSI-feedback bits.

InterDigital

**Proposal 3: TSF compression performance should be evaluated under multiple observation window lengths.**

**Proposal 4: For AI-TSF compression Case 2 with missing UCI, study different missing UCI mitigation solutions, e.g., buffer reset, to handle error propagation.**

NEC

***Proposal 8: At least for Case 2, CSI buffer reset should be supported to address*** ***misalignment of historical CSI used at UE side and NW side. And the definition, determination or indication of the reset value need to be further studied.***

***Proposal 9: At least for Case 2, further study*** ***effective availability of historical CSI information over time.***

***Proposal 10:*** ***If the*** ***CSI reconstruction part/model at UE side is available, support Option 2a for CQI determination.***

***Proposal 11: For defining the pairing information used to enable the UE to select a CSI generation model(s) that is compatible with the CSI reconstruction model(s) used by the gNB, down select from the following options:***

* ***Option 1: The pairing information is in the forms of the CSI reconstruction model ID that NW will use.***
* ***Option 2: The pairing information is in the forms of the CSI generation model ID that the UE will use.***
* ***Option 3: The pairing information is in the forms of the paired CSI generation model and CSI reconstruction model ID.***

Apple

**Proposal 6: For time-frequency-spatial domain CSI compression, flexible CSI report configuration to support different cases should be studied.**

Qualcomm Incorporated

1. Conclude that CQI calculation option 2a (where UE runs CSI reconstruction model and use its output for CQI calculation) can be employed with the consideration of potentially higher timeline and higher cost of processing unit and memory.
2. Further study CQI calculation option 1b (CQI is calculated based on target CSI with realistic channel measurement and potential adjustment) considering adjustment measurement at UE side based on intermediate KPI or intermediate output of the CSI generation model
3. Consider layer-common and rank common (Option 3-1) structure for CSI generation model and/or CSI reconstruction model for specified structures. Layer-common (Option 3-1) or layer-specific (Option 2-1) parameters can be upto vendor’s implementation choice.
4. Note: The standardized model structure is used to address inter-vendor collaboration complexity. The specification should be flexible to allow actual model for inference designed using all options 1-1, 1-2, 2-1, 2-2, 3-1 and 3-2.
5. Study following levels of quantization alignment from the aspect of scalability across vendors, performance, inter-vendor collaboration complexity

* Level 1: Proprietary quantization configuration and codebook, and exchange of both of them.
  + Note: the exchange can be via standardized signalling or proprietary signalling
* Level 2: standardization of quantization configuration and exchange of quantization codebook
  + Note: the exchange can be via standardized signalling or proprietary signalling

CEWiT

**Proposal-2: Study methods to model the absence of past CSI in the case of rank adaptation in Case-3 and Case-4 based CSI compression.**

**Proposal-8: Model pairing procedure to be performed before inference operation, with the assistance of UE capability report information to ensure NW sided model can avoid any model mismatch.**

**Proposal-9: In case of improving inter-vendor collaboration, store the additional information of an NW-sided model like vector-quantisation codebook name or its properties (size, feature length).**

**Proposal-20: For AI/ML based CSI compression, further study configurations and related aspects for various model options.**

**Proposal-21: For CQI determination, consider Option-1(*CQI is NOT calculated based on the output of CSI reconstruction part from the realistic channel estimation*) to be the starting point.**

**Proposal-22: For CQI determination, Option 1c can be deprioritised**

## Discussion

|  |  |
| --- | --- |
| Companies | Views |
| Futurewei | Support CQI option 2a if decoder is available at UE  Molel pairing discussion deferred |
| Huawei | Standardized quantization alignment   * VQ: configuration/reporting/updating of quantization dictionary, segmentation of encoder output * SQ: configuration of granularity/range   CQI option 1 as starting point  CSI priority rule, CPU, UCI mapping  For high rank model aspects, support 3-1 (layer-common, rank-common), 3-2 (layer-common, rank-specific), and 2-1 (layer-specific, rank-common). FFS their LCM as one model or multiple |
| Spreadtrum, BUPT | Support CQI 1b w/ potential adjustment  Report historical CSI to handle UCI missing |
| Google | Support channel / precoder / precoder in beam domain for CSI compression  ML CSF has higher priority than others  Support CPU enhancement with MPU for measurement and IPU for inference  Hybrid ML-CSF (if low rank) + non-AI CSF (if high rank) |
| ZTE | Prioritize CQI 1a (based on target CSI), 1b (based on target CSI + bias) |
| Xiaomi | Enhance existing priority rule to accommodate ML CSF  Division of CSI part 2 component into N groups  How to pack multiple CSI in one report for prediction case  Handling missing of historical CSI |
| Fujistu | Target CSI being precoder in SF domain or angle-delay domain  Study pairing procedure, UE or NW initiated, pairing information, etc  Study using local ID for alignment to reduce overhead compared to global  Model ID is sufficient, pairing ID is not needed  To reduce workload, downselect between 1) model-specific configuration / format, and 2) a configuration / format adapting various possibilities  Deprioritize CQI option 2  Include layer ordering information in CSI part 1  UCI omission / priority based on layer ordering (based on index or singular values) |
| CATT | Standardized quantization alignment  Consider CQI option 1, reuse CQI quantization |
| Panasonic | Deprioritize CQI 1c (based on legacy codebook) and 2b (via precoded CSI-RS) |
| TCL | Consider AIML dedicated resource config, report config, report quantity  Enhance CSI priority rules  UE capability description |
| LGE | Study mechanisms to manage the accumulated past CSI at two sides  Study format of historical CSI and how to report it for performance monitoring  Two-step to resolve UCI missing: 1) check perf degradation due to historical CSI missing, 2) report past CSI  For CQI, consider option 1 or 2a |
| Lenovo | Support procedures/signaling that enables model having both SQ and VQ for CSI feedback bits generation |
| InterDigital | Study methods of handling UCI missing |
| NEC | Study buffer reset to handle UCI missing  Study effective availability of past CSI over time  Support CQI 2a (based on reference decoder)  Pairing based on encoder ID, decoder ID or pairing ID |
| Apple | Flexible CSI configuration to accommodate various cases. |
| QC | Conclude CQI 2a is feasible if UE is able to run it with additional cost of processing unit and memory  Further study CQI option 1b.  Consider layer-common structure (3-1), parameters can be layer-common or layer specific, for the model structure in inter-vendor collaboration option 3. Actual choice is upto implementation.  Further study standardized or spec-transparent quantization alignment. |
| CEWiT | Study method handling UCI missing or misalignment  Consider CQI option 1a, 1b, deprioritize CQI option 1c. |

CQI options are discussed by 11 companies. CQI option 2a (based on reference decoder at UE side) is favoured by 5 companies. CQI option 1 is favoured by 9 (7 for 1a, 9 for 1b)

Model design aspects regarding high rank support is discussed by 2 companies, layer-common rank-common, layer-common rank-specific, layer-specific rank-common are favored.

Quantization alignment methods is discussed by 4 companies, 3 of them are interested in standardized approach, while 1 company mention it can be achieved via model / parmater / dataset sharing in inter-vnedor collaboration options. The configuration and indication of codebook / dictionary are also discussed.

UCI missing / misalignment handling for case 2 / 4 are discussed by 4 companies. Candidate schemes of buffer reset, historical CSI reporting mechanism are mentioned.

CSI priority rule, CSI processing criteria (CPU), UCI mapping are discussed by 5 companies. ML CSI has different (or higher) priority than other non-AI CSI. Different processing unit may be needed for measurement and inference. CSI part 2 may be needed to be divided into N part with different mapping / dropping priority. CSI in part 2 may be ordered by layer indices, the dropping may be based on index or singular values.

Model paring procedure is discussed by 2 companies regarding UE-initiated procedure vs. NW-initiated, model ID vs. pairing ID, global vs. local ID

Discussion 51a:

FL does not think that the above aspects are critical for determining feasibility of the CSI compression use case. They can generally be studied in the normative phase. Therefore, FL wants to deprioritize discussion of the above aspects.

If companies disagree with the FL’s assessment and believe that certain aspects are critical for determining feasibility of the CSI compression, please state those aspects and why.

|  |  |
| --- | --- |
| *Support / Can accept* |  |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
|  |  |
|  |  |

### Others

Please provide any other comments for this section.

|  |  |
| --- | --- |
| *Company* | *Comments* |
|  |  |
|  |  |

# Complexity performance trade-off

## Summary of company proposals

## Discussion

Regarding complexity, it is noted by companies that

* Use of localized models may reduce complexity.
* Use of temporal aspects may reduce complexity.
* Use of transformed domain input (e.g., angular, delay, Doppler) may reduce complexity.
* Knowledge distillation and pruning techniques may be utilized to reduce complexity with negligible performance loss.

Below, the FL captured the SGCS vs. complexity for Case 0, Case 2, and Case 3 based on companies’ submissions in the Results Spreadsheet. For now, these are provided for informational purposes. Later on, we may want to capture some observations from these.

It is important to note that many of the AI/ML models used in the evaluation study may not have been optimized in terms of complexity.

It is important to note that the actual AI/ML model complexity depends on various platform-dependent choices for modem implementation. The values reported here should be considered as representative values and not as a precise complexity estimate for implementation of AI/ML-based CSI compression.

### Performance vs. complexity for temporal domain Case 0

In all the plots, the x-axis is the combined complexity of the CSI generation part and the CSI reconstruction part.

### Performance vs. complexity for temporal domain Case 2

In all the plots, the x-axis is the combined complexity of the CSI generation part and the CSI reconstruction part.

### Performance vs. complexity for temporal domain Case 3

In all the plots, the x-axis is the combined complexity of the CSI generation part and the CSI reconstruction part.

For Case 3, the FLOPs represent the total FLOP over the time window, not the normalized FLOP over 5msec. (i.e., the FLOPs is based on FLOPs/M from the results template, not the FLOPs/M/5msec.)

### Others

Please provide any other comments for this section.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Intel | Thanks to the FL for summarizing the model complexity in the figures.  Is it planned to capture the figures in the TR or as part of an observation? |
|  |  |

# Other topics (work plan, new use cases)

## Summary of company proposals

From the submitted contributions, proposals related to other aspects not covered in other sections (work plan, new use cases, etc.) are summarized below.

Huawei

***Observation 9: Compared with Rel-18 SF domain CSI compression, involving additional temporal domain compression may bring additional performance gains from the evaluation of the study. On the other hand, the potential additional spec impact of the temporal domain cases has not been sufficiently assessed to justify the feasibility, especially for the inter-vendor training collaboration aspects, e.g., TSF specific model structure and dataset format.***

***Observation 10: RAN1 may need more time/effort to justify the feasibility for standardizing Option 1/3/4/5, considering:***

* ***For Option 1/3, the discussions leading to an exact model/model structure to be standardized have not started and are expected to be time consuming.***
* ***For Option 4/5, the discussions of specific dataset format and model representation format have not started.***
* ***For Option 3/4/5, the specific path (over-the-air or other approaches) and corresponding solutions for how to deliver parameter/dataset/model have not been sufficiently investigated.***
* ***Further down selection among options and sub-options may be needed to limit the scope of the study.***

***Observation 13: RAN1 may need more time/effort to justify the feasibility for some essential Rel-18 leftover issues, including:***

* ***For data collection, NW side data collection type/format/signalling have not been analysed, without which it is unlikely to support parameter/dataset/model delivery from NW side to UE side (i.e., NW side train/NW first train).***
* ***For monitoring, monitoring options have been discussed, yet more discussion is needed to identify the feasibility of monitoring for both sides.***

OPPO

***Recommendation: suggest to proceed into WI phase for AI/ML based CSI compression in Rel-19 after RAN#105 in September:***

* ***At least further study AI/ML based CSI compression in Rel-19, if WI is not established.***

Intel

***Observation 6***:

* *High model complexity, limited performance gains, unresolved inter-vendor training collaboration, and testability issues make it very challenging to support AI/ML CSI compression with two-sided model in Rel-19 timeframe.*

Xiaomi

***Proposal 14: Recommend the two-sided AI/ML model based CSI compression to study as a normative work, and at least Case 2 and Case 3 should be supported.***

Fujistu

***Proposal 3:***

* *Regarding the use cases for CSI feedback enhancement studied in Rel-19, if AI/ML based CSI compression with two-sided model will be specified in Rel-19, it is recommended to consider Case-2 and Case-3 for normative work based on the evaluation results*
  + *Case-2 could be considered for indoor scenario*
  + *Case-3 could be considered for both indoor and outdoor scenario*
* *If the time is not sufficient to specify AI/ML based CSI compression within Rel-19 timeframe, further evaluation and study could be considered in Rel-19.*

***Proposal 25:***

* *If two-sided model for CSI compression will be specified in Rel-19, Option 4 for alleviating/resolving the inter-vendor collaboration issues is recommended for normative work.*

InterDigital

**Proposal 5: AI/ML CSI compression should continue as a study item for the remainder of Rel-19.**

NTT Docomo

**Observation 5**

* **During the Rel. 19 study, the performance gain of AI/ML-based CSI compression has been further improved on top of the achievements of the Rel. 18 study with techniques such as CSI compression with temporal domain aspects.**

**Observation 6**

* **On inter-vendor collaborations,**
  + **RAN1 and other RAN WGs, such as RAN4, have identified options to solve the inter-vendor collaboration and interoperability/testability issues. Among them, significant progress has been achieved on Option 1.**
    - **Option 1 can serve as a baseline scheme to solve the inter-operability and testability with the most minor complexity, i.e., the simplest way, based on the state of the art.**
    - **Although the flexibility of Option 1 is limited, there is still room for Option 1 to adapt the environments for performance in real-life deployments.**

**Proposal 4**

* **RAN1 recommends the normative work or further study on inter-vendor collaboration Option 1 for Rel-19.**
  + - **Other options are not precluded if their feasibility can be justified.**

Intel

***Observation 1***:

* *For the case of 32 CSI-RS ports, 2 Rx antennas at the UE, 52 PRB bandwidth, 4 PRB subband size, CSI-RS density 1, Enhanced Type II PMI codebook with L = 4, N3 = 13, M = 4, K0 = 16.*
  + *Complexity of PMI search is ~1 MFLOPs.*
  + *Complexity of PMI reconstruction is ~0.02 MFLOPs.*

***Proposal 1***:

* *For the observations with AI/ML CSI compression performance gains over PMI codebook, capture range for the AI/ML model complexity together with the corresponding range for performance gains.*
  + *For all the observations with AI/ML model complexity, the following note should be added: “AI/ML model complexity depends on various platform-dependent choices for model implementation. The values reported here should be considered as representative values and not as a precise complexity estimate for implementation of AI/ML-based CSI compression.”.*

Vivo

1. **By considering TSF compression studied in R19, two-sided model complexity could be reduced compared with that of models in R18 SF compression. Our simulation results reveal that the parameter scale/FLOPs can be reduced to ~1/10 while achieving even higher performance gain.**

CATT

Observation 4: There are many state-of-the-art techniques in Machine Learning for complexity reduction that can be applied to AI/ML based CSI compression. These approaches are implementation specific and have little specification impact.

## Discussion

### Others

Please provide any other comments for this section.

|  |  |
| --- | --- |
| *Company* | *Comments* |
|  |  |
|  |  |

# Proposals for online sessions

## Proposals for Monday online session

## Proposals for Tuesday online session

## Proposals for Wednesday online session

## Proposals for Thursday online session

## Proposals for Friday online session

# FL closing remark

# List of agreements

## Agreements from RAN1 #116

**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 |

Note 1: For the UE, the past CSI information may include past model inputs and/or any information derived from them. For the network, the past CSI information may include past CSI feedback instances and/or any information derived from them.

Note 2: For case 3 and case 4, the UE may perform prediction as a separate step or jointly with compression. Similarly, the network may perform prediction as a separate step or jointly with reconstruction. Companies to report which option is selected, the number of future slots, and whether the prediction is AI/ML-based or not.

Note 3: “Target CSI slot(s)” refers to the slot(s) to which the CSI feedback in the report corresponds. “Present slot” refers to the slot of the most recent CSI-RS measurement used to generate the CSI report. “Future slot(s)” includes at least one slot after the present slot and may include the present slot as well.

Note 4: Down-selection is not precluded.

**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.

**Working Assumption**

For the evaluation of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Release 19, adopt the following benchmark scheme for performance comparison:

* For cases without prediction of future CSI, use the same benchmark scheme assumed in R18 AI/ML-based CSI compression study.
* For cases with prediction of future CSI, use the same benchmark scheme assumed in R18 AI/ML-based CSI prediction study, with R18 MIMO eType II codebook for compressing the feedback.

**Agreement**

For the evaluation of AI/ML-based CSI compression using localized models in Release 19, study the following aspects of the performance/complexity trade-off when comparing the localized model with a benchmark model that is not localized:

* Performance of the localized model that has similar or lower complexity as the benchmark model.
* Model complexity of the localized model that achieves similar or better performance as the benchmark model.

**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**

To alleviate / resolve the issues related to inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model, study the following options:

* Option 1: Fully standardized reference model (structure + parameters)
* Option 2: Standardized dataset
* Option 3: Standardized reference model structure + Parameter exchange between NW-side and UE-side
* Option 4: Standardized data / dataset format + Dataset exchange between NW-side and UE-side
* Option 5: Standardized model format + Reference model exchange between NW-side and UE-side

Note 1: The above options may not be mutually exclusive and may be used together.

Note 2: Other options are not precluded.

Note 3: The study should consider how different methods of exchanging the parameters / dataset / reference model would affect the feasibility and collaboration complexity of options 3 / 4 / 5 respectively, e.g., over the air-interface, offline delivery, etc.

Note 4: “Dataset” refers to a set of data samples of CSI feedback and associated target CSI.

**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.

**Agreement**

* For the evaluation of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Release 19,
  + adopt the CSI feedback overhead rate as reference, where the CSI feedback overhead rate is the average bit-rate of CSI feedback overhead across time.

Note: The CSI feedback overhead of a single report is calculated as in R18 CSI compression study.

**Agreement**

For the evaluation of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Release 19, for cases with prediction of future CSI, in which prediction and compression are separated, to optionally evaluate a scheme with ideal prediction as an additional evaluation case for reference.

Note: The ideal prediction scheme should model realistic channel estimation.

**Agreement**

For the evaluation of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Release 19, for Case 2, Case 4 and Case 5, study the performance impact resulting from non-ideal UCI feedback.

**Agreement**

For the study of inter-vendor collaboration issues for AI/ML-based CSI compression using a two-sided model, consider at least the following aspects when comparing different options:

* Inter-vendor collaboration complexity, e.g., whether bilateral collaboration is required between vendors.
* Performance.
* Interoperability and RAN4 / testing related aspects.
* Feasibility.

## Agreements from RAN1 #116-bis

Agreement

For the results template used to collect evaluation results for temporal domain compression Case 1/2/5, adopt Table 1 used in Rel-18 as starting point with the following additions:

* Temporal domain CSI setting
  + CSI feedback periodicity
  + CSI-RS periodicity
* Description of model input/output and Case
  + Compression case, e.g., Case 1/2/5
  + Usage of historical CSI at UE/NW side (e.g., number / time distance, eigen-vectors / raw channels, etc)
  + Methods to handle UCI loss (if applicable), e.g., CSI buffer reset, CSI retransmission, etc.
  + Methods to handle rank adaptation (if applicable)
* UE distribution (Option 1 or Option 2) and UE speed
* CSI feedback overhead rate: X/Y/Z bits per normalized time unit
  + Normalized time unit = 5ms and adopt same X/Y/Z values as in Table 1 of Rel-18
* Benchmark scheme
  + Rel-16 eT2 and compression Case 0 (i.e., Rel-18 AI/ML based CSI compression)
* Whether/how spatial consistency is modelled
* Whether/how UCI loss is modelled
  + The same UCI loss model shall be applied to the benchmark for fair comparison.
* Whether/how rank adaptation is modelled
* Modelling of channel estimation error
* Whether/how phase discontinuity is modelled (if applicable)

Agreement

For the results template used to collect evaluation results for temporal domain prediction and compression Case 3/4, adopt Table 1 used in Rel-18 as starting point with the following additions:

* Temporal domain CSI setting
  + CSI feedback periodicity
  + CSI-RS periodicity
* Description of model input/output and use case
  + Compression case, e.g., case 3 / 4
  + Observation window (usage of historical CSI at UE/NW side, e.g., number / time distance, eigen-vectors / raw channels, etc)
  + Prediction window (e.g., time distance between 1st prediction instance and last observation instance, number / time distance of predicted CSI)
  + Methods to handle UCI loss (if applicable)
* UE distribution (Option 1 or Option 2) and UE speed
* CSI feedback overhead rate: X/Y/Z bits per normalized time unit
  + Normalized time unit = 5ms and adopt same X/Y/Z values as in Table 1 of Rel-18
* SGCS values before (if applicable) and after compression
* Assumption on the prediction of future CSI
  + Separate step or jointly with compression
  + If separate, description of the AI or non-AI prediction algorithms: ideal prediction, AI-based prediction, non-AI-based prediction (e.g., nearest historical CSI and its location, learning window size / time correlation matrix size for auto-regression based prediction),
    - Note: the same prediction algorithm to be used for the benchmark scheme.
* Benchmark schemes
  + Description of feedback schemes, i.e., Rel-18 doppler eT2
* Whether/how spatial consistency is modelied
* Whether/how UCI loss is modelled
  + The same UCI loss model shall be applied to the benchmark for fair comparison.
* Modelling of channel estimation error
* Whether/how phase discontinuity is modelled (if applicable) ~~Modelling of phase discontinuity~~

Conclusion

For multi-vendor results table, adopt Rel-18 Table 4 for joint training and Rel-18 Table 5 for separate training as starting point, with the same additions of above 2 agreements.

Conclusion

For model generalization results table, adopt Rel-18 Table 2 and Generalization Case 1 / 2 / 3 as starting point with same additions above. For generalization aspects, adopt the following

* Various UE speed
* UE distribution
* Various CSI-RS periodicity

Conclusion

For model scalability results table, adopt Rel-18 Table 3 and Generalization Case 1 / 2 / 3 as starting point with same additions above. For generalization aspects, adopt the following

* Various numbers of antenna ports
* Various frequency granularity
* Various payload size

Conclusion:

* Conclude, from RAN1 perspective, that Option 1, if feasible for specification, eliminate the inter-vendor collaboration complexity (e.g., whether bilateral collaboration is required between vendors).
* It is RAN1’s understanding that Option 1 corresponds to RAN4 options, e.g., RAN4-Option3, or RAN4-Option4. Further study and final conclusion on interoperability and RAN4 testing of the RAN4-Option3 and RAN4-Option4 is up to RAN4.

Observation

* Option 1 and 2 may have limited performance in the field compared to Options 3, 4, and 5, further study is needed
* Option 1 and 2 may require high specification effort from RAN1 perspective.

Conclusion

* Deprioritize Option 2 for inter-vendor training collaboration.
  + Note: This deprioritization shall not affect the ongoing discussion in RAN4 on RAN4-Option3 and RAN4-Option4.

Agreement

* 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.
* Note: The descriptions under each option are only for the purpose of simplified discussion and do not mean deprioritizing any other flavors (such as an exchange originating from the UE-side and ending at the NW-side) from potential specification.

Agreement

* 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.
  + Note: For each option/sub-option of interest, companies to bring discussion on how inter-vendor collaboration complexity, interoperability, and feasibility may be addressed. Companies to strive to provide solution(s) that can address all the following aspects: inter-vendor collaboration complexity, performance, interoperability, and feasibility.
  + Note: The descriptions under each option are only for the purpose of simplified discussion and do not mean deprioritizing any other flavors (such as an exchange originating from the UE-side and ending at the NW-side) from potential specification.

Agreement

* For the results template used to collect evaluation results for AI/ML-based CSI compression using localized models, adopt Table 1 used in Rel-18 as starting point, capturing the generalized model result and the localized model result as separate columns, with the following additions for the localized model:
* Dataset description
  + Local region modelling: e.g., Option 1 or Option 2, and further details
  + Temporal modelling: e.g., how temporal variation is modelled in train and test sets
  + Dataset description for generalized model

Conclusion

In Rel-19 study of temporal domain aspects of AI/ML-based CSI compression using two-sided model, CSI prediction that is performed entirely at NW-side is deprioritized.

Agreement

* For the evaluation of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Release 19, for the temporal domain prediction and compression Case 3 and Case 4, adopt the following evaluation assumptions as baseline:
  + Observation window (number**/**distance):
    - For periodic CSI-RS with 5ms periodicity: 12/5ms, 10/5ms, 8/5ms, 5/5ms, 4/5ms, unrestricted observation window
    - For periodic CSI-RS with 20ms periodicity: up to companies (encouraged)
    - For aperiodic CSI-RS: 12/2ms, 8/2ms, 4/2ms
    - Others can be additionally submitted
  + Prediction window (number**/**distance between prediction instances**/**distance from the last observation instance to the 1st prediction instance): 4/5ms/5ms
    - Others can be additionally submitted, e.g. 4/1ms/5ms, 8/1ms/5ms, 4/5ms/10ms, 1/-/5ms

Agreement

For the results template used to collect evaluation results for temporal domain prediction and compression Case 4, adopt Table 1 used in Rel-18 as starting point with the following additions:

* Description of model input/output and use case
  + Methods to handle rank adaptation (if applicable)

## Agreements from RAN1 #117

Conclusion

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.
* Standarized 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.
* Standarized signalling may be over-the-air, or other approaches.

Note: feasibility will be discussed separately.

Agreement

* 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.

Agreement

For option 1 / 3 / 4 / 5 and their sub-options, study mechanisms (e.g., post-deployment performance monitoring) for identifying the cause (e.g., NW side, UE side, data drift) of the performance degradation to guarantee good performance in the field.

Agreement

For temporal domain aspects Case 3/4, change the small / medium / large payload region definition as follows:

|  |
| --- |
| Note: X, Y, Z, A, B, and C are feedback overhead rates in bits per time unit of 5ms.  Note: For X, Y, and Z, α=[2] for rank=1/2 and α=[4] for rank=4  Note: For A, B, and C, β=[0.5] for rank=1 and β=[0.75] for rank=2/4 |

Agreement

For the evaluation of temporal domain aspects of AI/ML-based CSI compression (Cases 1-5), in addition to FLOPs, also consider FLOPs per normalized time unit. Use 5msec as the normalized time unit.

Agreement

In the results template for capturing the evaluation of temporal domain aspects Case 3/4 of AI/ML based CSI compression, regarding the “upper bound”, capture both of the following:

* upper bound based on ideal CSI prediction and without CSI compression
* upper bound based on benchmark CSI prediction and without CSI compression

Agreement

For the evaluation of AI/ML-based CSI compression using localized models in Release 19, regarding training,

* The k-th local model is trained on region #B\_k (the k-th local region), 1<=k<=N.
* The generalized model is trained on Region #A that may be constructed via any of the following methods that is appropriate for the given generalized/local region modeling approach.
  + Region #A is the same as the union of regions #B\_1, …, #B\_N.
  + Region #A is a proper superset of the union of regions #B\_1, …, #B\_N.
  + Region #A is generated separately from regions #B\_1, …, #B\_N.
  + Note: companies to report which method was used.

For the evaluation of AI/ML-based CSI compression using localized models in Release 19, regarding testing,

* The trained generalized model, local model, and the non-AI/ML benchmark are tested on the regions #B\_1, …, #B\_N.
* In case N>1, when reporting the results, companies may report the performance of the generalized model, the local models, and the non-AI/ML benchmark, by averaging the performance over the regions #B\_1,…,B\_N. Companies to report the value of N.

Agreement

For collecting evaluation results for temporal domain aspects of AI/ML-based CSI compression using localized models, use the same results template used to collect evaluation results for AI/ML-based CSI compression using localized models

* Adding the same temporal setting that is used for results template used to collect evaluation results for temporal domain compression Case 1/2/5.

|  |  |
| --- | --- |
| **Temporal setting** | **Temporal domain aspect Case 1-5** |
| **CSI-RS configuration: periodic or aperiodic For periodic: periodicity For aperiodic: # of resources K in the CSI-RS burst / time internal m in msec** |
| **CSI reporting periodicity** |
| **Usage of historical CSI at UE side: number / time distance** |
| **Usage of historical CSI at NW side: number / time distance** |
| **Prediction window: number / time distance between prediction instances / distance from the last observation instance to the 1st prediction instance (Only applicable to Case 3,4)** |

Agreement

Further study following monitoring options in Rel-19, including the necessity and feasibility,

* NW-side monitoring, considering overhead, latency, complexity, monitoring accuracy, UE capability
  + Based on the target CSI reported by the UE via legacy eT2 codebook or eT2-like high-resolution codebook (Case 1)
  + SRS-based monitoring
* UE-side monitoring, considering overhead, latency, complexity, monitoring accuracy, UE capability
  + Based on the output of the CSI reconstruction model at the UE (Case 2-1)
    - Note: CSI reconstruction model at the UE-side can be the same as the actual CSI reconstruction model used at the NW-side, a reference model provided by NW, or a proxy model developed by the UE side.
  + Via direct estimation of intermediate KPI (e.g., SGCS) without reconstructing a target CSI (Case 2-2)
  + Via estimation of monitoring output other than intermediate KPI without reconstructing a target CSI
  + Based on precoded RS (e.g., CSI-RS, DMRS) transmitted from NW based on the output of the CSI reconstruction model
  + Based on the output of the CSI reconstruction model indicated by the NW via legacy eT2 codebook or eT2-like high-resolution codebook

Regarding monitoring metrics:

* Monitoring accuracy also includes generalization considerations, if applicable.
* Complexity also includes LCM complexity, if applicable.
* Monitoring overhead, latency, complexity, and accuracy analysis may have to consider using at N>1 CSI feedback occasions.
* Testability of UE reported metrics

Discussion may include the following aspects:

* Consideration of Options 1-5 and their sub-options for alleviating / resolving the issues related to inter-vendor training collaboration
* Temporal domain aspects of CSI compression
* How the above monitoring approaches or combination of them may help identifying the cause (e.g., NW side, UE side, data drift) of the performance degradation

Note: for UE-side monitoring, the final reported monitoring output, if specified, may be different, e.g., be further derived based on the output of the above approaches.

Note: implementation-based monitoring solutions can be considered in assessing the necessity of the above monitoring approaches.

Agreement

For temporal domain aspects Case 3 and 4, study the impact on LCM aspects of separate prediction and compression, and joint prediction and compression.

Note: Observations of companies results till RAN1#117 are captured in FL summary R1-2405419.

## Agreements from RAN1 #118

# References

1. TR 38.843 v18.0.0, “Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface” (Release 18), December 2023.
2. RP-234039, “New WID on Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface”, Qualcomm (Moderator), 3GPP TSG RAN #102, December 2023.
3. R1-2401561, “Revised Final summary for Additional study on AI/ML for NR air interface: CSI compression”, Moderator (Qualcomm), 3GPP TSG RAN WG1 #116, Feb. 2024.
4. R1-2403505, “Final summary of Additional study on AI/ML for NR air interface: CSI compression”, Moderator (Qualcomm), 3GPP TSG RAN WG1 #116-bis, Apr. 2024.
5. R1- 2405419, “Final summary of Additional study on AI/ML for NR air interface: CSI compression”, Moderator (Qualcomm), 3GPP TSG RAN WG1 #117, May, 2024