**3GPP TSG-RAN WG1 Meeting #117 R1-** **240xxxx**

**Fukuoka, Japan, May 20th-24th, 2024**

**Agenda item:** 9.1.3.2

**Source:** Moderator (Qualcomm)

**Title:** 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#116-bis.

## 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 | yamamoto.tetsuya001@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 |

# 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

FUTUREWEI

***Proposal 2: For the evaluation of temporal-domain CSI compression Case 1, consider at least one implementation option that the compressed representation of CSI generation part output may contain the past CSI information.***

Tejas Networks Ltd.

***Proposal 1: For the EVM of temporal domain CSI compression Case 2, consider the following assumptions for the CSI generation part and CSI reconstruction part, respectively:***

* ***CSI generation part at t=T:***
* ***Model input: Predicted CSI for the current slot*** ***() (it is not necessarily needed to be from the current slot) () from the previous time slots at UE.***
* ***Model output: CSI feedback of the current slot () can be passed to AI-based quantizer to get quantized bit of sequence ().***
* ***CSI reconstruction part at t=T+∂ (where ∂ is an uplink latency)***
* ***Model input: CSI feedback () of the current slot at NW will be passed to AI-based d-quantizer to get the floating-point values which will again be passed to decoder along with accumulated CSI information () from the last time instance (T-1+∂)***
* ***Model output: Recovery CSI of the current slot () and accumulated CSI information for the next time instance ().***

**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.***

Huawei, HiSilicon

***Proposal 1: For the evaluation of temporal domain based CSI compression Case 1~Case 5, take Case 2 and Case 3 with higher priority for evaluation.***

***Proposal 3: 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 4: 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.***

***Proposal 5: For the UE distribution EVM assumption, consider 80% indoor, 20% outdoor for temporal domain CSI compression cases without future CSI, and consider 100% outdoor for temporal domain CSI compression cases with future CSI.***

***Proposal 6: For the particular potential spec impact to support temporal CSI compression Case 3, it may need to be discussed whether these two features are regarded as one model or two separate models from the LCM perspective.***

* ***The impacted LCM procedures include, e.g., data collection, monitoring, inference, model control (activation/deactivation/switching/fallback), etc.***

Intel Corporation

***Proposal 1***:

* *PMI search and PMI reconstruction complexity assumed for performance evaluation of eType II PMI codebook should be disclosed by companies.*
  + *Alignment of PMI search and PMI reconstruction complexity assumption should be further discussed in RAN1.*
  + *If CSI with PMI prediction (Enhanced Type II codebook for predicted PMI) is used as benchmark for CSI compression, complexity of utilized prediction algorithm shall be disclosed by companies.*

***Proposal 2***:

* *The following aspects are also considered for the additional study on AI/ML CSI compression:*
  + *Complexity of CQI determination.*
  + *UE and gNB memory size requirements.*

InterDigital

**Proposal 1: Increased complexity of TSF over SF should be taken into account when investigating the potential benefit of TSF.**

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

Samsung

**Proposal#1: Among the identified six categories 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 categories 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 categories 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#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**

Apple

**Proposal 1: For time-frequency-spatial domain CSI compression, the following potential specification impact are proposed:**

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

**Proposal 2: For time-frequency-spatial domain CSI compression, flexible CSI report configuration to support different cases should be studied.**

CATT

Proposal 1: For temporal domain aspects of AI/ML-based CSI compression using two-sided model in Rel-19, if current CSI slot is targeted, Case 2 is prioritized for study.

Proposal 2: For temporal domain aspects of AI/ML-based CSI compression using two-sided model in Rel-19, separate CSI prediction and CSI compression model is prioritized for Case 3 and 4.

**Proposal 3: For temporal domain aspects of AI/ML-based CSI compression using two-sided model in Rel-19, considering the following sub-cases for Case 3:**

* **Case 3-1: target CSI is the CSI for a future instant**
  + **Case 3-1a: CSI generation part without past CSI information (CSI prediction+ Case 0)**
  + **Case 3-1b: CSI generation part with past CSI information (CSI prediction+ Case 1)**
* **Case 3-2: target CSI is the CSI for multiple future instants**
  + **Case 3-2a: CSI generation part without past CSI information (CSI prediction+ Case 0 multiple slots extension)**
  + **Case 3-2b: CSI generation part with past CSI information (CSI prediction+ Case 1 multiple slots extension)**

**Proposal 4: For temporal domain aspects of AI/ML-based CSI compression using two-sided model in Rel-19, if multiple future CSI slots are targeted, Case 3-2a which includes a direct multiple slots extension for the Rel-18 AI CSI compression should be prioritized for study.**

Proposal 5: For temporal domain aspects of AI/ML-based CSI compression using two-sided model in Rel-19, if future CSI slot(s) is (are) targeted, focus on cases with CSI compression part having a major impact on the overall performance. The CSI prediction should not be the bottleneck and ideal CSI prediction can be assumed.

Lenovo

**Proposal 19: Prioritize Case 2 and Case 3 for temporal domain aspects of AI/ML-based CSI compression using two-sided model.**

**Proposal 20: Strive to unify the CSI framework across the two agendas for AI/ML study of CSI feedback compression and CSI prediction enhancements.**

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 #3: 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.**

**Proposal #4: Consider the method on the rank adaptation based on the availability check of layer(s) for a given RI.**

SK Telecom

Proposal 2 For AI/ML-based CSI compression study using two-sided model in Release 19, temporal domain prediction and compression Case 3 and 4 (i.e., Target CSI slot(s) = Future slot (s)) are prioritized than others.

Proposal 3 Consider to study combined P-CSI-RS + AP CSI-RS for AI/ML-based CSI compression, if it is concluded that there is no meaningful gain of AI/ML based CSI compression with 20ms p-CSI-RS only.

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.***

NEC

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

***Proposal 2: For Case 2, CSI buffer reset should be supported to address UCI loss and rank adaption. And the definition, determination or indication of the reset value need to be further studied***

***Proposal 3: For Case 2, further study*** ***effective availability of past CSI information over time.***

***Proposal 4: Study to use model compression to reduce the complexity of the AI/ML model at least for Case 2.***

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

***Proposal 3: For results template Table 1, further study the potential new boundary values to capture the evaluation results of high feedback overhead for temporal domain CSI compression Case 3/4 when the length of prediction window is larger than 2.***

***Proposal 4: For* *temporal domain CSI compression Case 2, further study the performance impact resulting from the aperiodic CSI feedback.***

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.

Proposal 5: For AI/ML-based CSI compression using two-sided model, when the target CSI is Future slot(s), study method to align whether prediction and compression occur in separate steps or simultaneously between UE and NW, for inference and performance monitoring operations.

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*

Nokia

Proposal 1: Regarding SFT-based CSI compression, prioritize Case 2 and Case 4, as these cases are more promising in terms of delivering significant gains.

Proposal 2: For SFT based CSI compression, time coherence effect should be studied, e.g., different scenarios, different UE speeds, various CSI feedback periods, and different length of time sequences.

Proposal 3: Non-ideal UCI feedback impact should be studied and monitoring mechanisms that detect such non-ideal feedback should be developed. Additionally, mechanisms to re-synchronize historical CSI information should be studied.

AT&T

**Proposal 1: Joint CSI compression and prediction is deprioritized for Rel-19.**

**Proposal 2: For CSI compression and prediction, only consider the scenarios where only the UE is performing the CSI prediction in Rel-19.**

MediaTek Inc.

Proposal 1. Evaluate effectiveness of complexity reduction techniques in reducing both computational and storage complexities of AI/ML models for CSI compression

Proposal 2. Consider and evaluate eType II algorithm’s complexity as the baseline of computational complexity.

Proposal 3. Evaluate the feedback error tolerance of eType II and compare it with that of AI/ML model.

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

### Overhead range adjustments

Currently, for temporal aspects of CSI compression, the small / medium / large payload regions are define 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=1 for rank=1/2 and α=2 for rank=4  Note: For A, B, and C, β=1 for rank=1 and β=1.5 for rank=2/4 |

However, for temporal domain aspects Case 3/4, given the max payload size of Rel-18 eTypeII and normalization per 5msec, and given the better compression via temporal domain achieved by Case 3/4, we may want to adjust the range of X/Y/Z and A/B/C. In fact, ZTE noted that, with prediction window length of 4, even the highest parameter combination level (i.e., PC9) of Rel-18 eTypeII will belong to region Y and therefore region Z will be empty.

|  |
| --- |
| To our understanding, the highest parameter combination level (i.e., PC9) of Rel-18 eTypeII is less than 700 bits under Rank=1. If prediction window length is 4, the feedback overhead rate is less than 175 bits/instance, which cannot satisfy the feedback overhead Z(i.e., >=230 bits). So, we would like to know how to resolve this situation. |

This could be done by either directly changing the boundary values 80, 100, 140, 230, or by changing the value of α and β. The FL is prosing the second approach:

Proposal 1a:

For temporal domain aspecs 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 |

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Fujitsu | Generally fine with the proposal. |
|  |  |

### Model complexity FLOPs

Discussion 2a:

For the evaluation of temporal domain aspects of AI/ML-based CSI compression (Cases 1-5), given that we are using the CSI feedback overhead rate normalized by 5msec for fair comparison among schemes of different feedback periodicities, should we make similar changes in capturing the model complexity FLOPs/M? For example, should we use the average FLOPs/M/5msec instead of FLOPs/M for fair comparison between schemes of different periodicities? Please provide your suggestion.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | We are fine with this direction. Similar as overhead rate, we can also normalize the FLOPs/M by 5msec. |
| Huawei, HiSilicon | FLOPs reflects not only the average calculation complexity, but also the peak calculation complexity. Therefore, if an average FLOPs needs to be introduced, we think we should not only adopt FLOPs/5msec, but also capture the R18 FLOPs. |
| Samsung | We have the same view as Huwaei (Peak FLOP is the main performance factor). However, companies may additionally report the averaged FLOPs in case the FLOPs vary through the reports. |
| ETRI | CSI feedback rate is to ensure a fair comparison when evaluating the performance of methods with different feedback intervals. This is because the CSI feedback sizes directly performances. However, the computational complexity of AI/ML models (i.e., FLOPs) serves as a reference for understanding/comparing the performance of the methods and was not aligned among methods even before the introduction of different feedback intervals. Therefore, we believe that the FLOPs does not need to be normalized. |
| ZTE | We are not clear for average FLOPs/M/5ms, which needs further clarification. For example, if we adopt Case 3 with 4 predicted CSI instances, each of which follows 5ms periodicity, does it mean the whole model complexity (FLOPs/M) needs to be divided by 4? |
| Fujitsu | We prefer to use average FLOPs/M/5msec to make a fair comparison among the cases. |
| Intel | In our view the complexity for one CSI instance is important due to latency requirements for CSI feedback. Due to low CSI processing time the peak computational power (FLOPS per second) shall be large to accommodate large task (in FLOPs) in few millisecond.  So, complexity for one CSI instance (in FLOPs) shall be reported together with CSI complexity averaged over a period of time (in FLOPS/M/5msec). |
| OPPO | We also share similar view to Huawei and Samsung. |
| CATT | Ok with what Huawei and Samsung proposed. |

### Upper bound for Case 3/4

The results templates contain “Gain for upper bound without CSI compression over Benchmark”. As temporal aspects Case 3/4 also involves prediction, there is a question of whether the upper bound should be based on ideal prediction or realistic prediction. The FL thinks that the use of ideal prediction makes more sense.

Proposal 3a:

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.

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Support |
| Samsung | Ok. |
| ETRI | Support |
| ZTE | Fine with the proposal. |
| Fujitsu | We don’t see the necessity for this proposal.  The important thing is to compare the performance with the benchmark, not to compare with the upper bound. |
| LG | Fine |
| OPPO | We are OK. From our undersranding, the SGCS of ideal CSI prediction without CSI compression shoud be 1? |
| CATT | ok |

### Separate vs. joint prediction

For temporal domain aspects Case 3/4, 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.

Proposal 4a:

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.

|  |  |
| --- | --- |
| *Support / Can accept* | Futurewei, Huawei, HiSilicon, Samsung, ETRI, ZTE, LG |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | LCM aspects can be further studied after we have more evaluation results and observations on Case3 and 4. |
| ETRI | We support the proposal. We see the difference between the seperated and the jointly performed cases at least in performance monitoring procedure. |
| ZTE | Fine with the proposal. |
| Fujitsu | The definition of separate/joint prediction seems not very clear and should be clarified. For two physically separate models (one for CSI prediction and the other for CSI compression) but jointly trained, should this case be treated as separate prediction and compression or joint prediction and compression? |
| OPPO | Okay, further study on LCM aspects can be later after more evaluations on Case 3/4. |

### 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*, till RAN1 #117,

* For Layer 1,
  + 1 source [OPPO] observes performance gain of 4.9% 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 1** of AI/ML based CSI compression compared to the CSI compression Case 0 *~~benchmark~~ in terms of SGCS*, till RAN1 #117,

* For Layer 1,
  + 1 source [Futurewei] observes performance gain of 6.77-12.1% at CSI payload X (small payload)
  + 1 source [Futurewei] observes performance gain of 6.16-6.95% at CSI payload Y (medium payload)
  + Performance gain at CSI payload Z (large payload) is TBD

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/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* |
| vivo | Need more time to check results. Would provide comments later. |
| Futurewei | For small CSI payload X, our results show SGCS gain for Case 1 over Case 0 varies depending on number of past CSIs, quantization method, and code size and quantization level combination. Overall range is from -0.36% to 12.1%. The worst gain (-0.36%) was observed when applying SQ with code size = 8 and 5bits/code while the highest gain was observed when integrating 3 past CSIs and applying VQ/LUT with 14 CSI overhead bits. |
| ZTE | To our understanding, it is too early to draw an observation for the case that only one source provides the simulation results, which can not reflect the actual range of gain for the case. So, we suggest waiting for more companions’ evaluation results to draw an observation. |
| Intel | In the current release we are looking into the tradeoff between perfoamance and complexity/overhead. So, we think that range of AI/ML model complexities shall be captured in observations (applies to all the observation with performance gain and complexity saving). |

Observation 112a: FTP traffic 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 #117,

For Max Rank 1, performance gain is TBD.

For Max Rank 2:

* For RU <= 39%, performance gain is TBD.
* For RU 40-69%, performance gain is TBD
* For RU > 70%,
  + 1 source [Futurewei] shows performance gain of 26% at CSI feedback overhead A (small overhead)
  + Perormance gain at CSI feedback overhead B (medium overhead) and C (large overhead) is TBD.

For Max Rank 4, performance gain is TBD.

For the evaluation of temporal domain aspects **Case 1** of AI/ML based CSI compression compared to the CSI compression Case 0 *~~benchmark~~ in terms of mean UPT* *under FTP* traffic, till RAN1 #117,

For Max Rank 1, performance gain is TBD.

For Max Rank 2:

* For RU <= 39%, performance gain is TBD.
* For RU 40-69%, performance gain is TBD
* For RU > 70%,
  + 1 source [Futurewei] shows performance gain of 2% at CSI feedback overhead A (small overhead)
  + Perormance gain at CSI feedback overhead B (medium overhead) and C (large overhead) is TBD.

For Max Rank 4, performance gain is 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 #117,

For Max Rank 1, performance gain is TBD.

For Max Rank 2:

* For RU <= 39%, performance gain is TBD.
* For RU 40-69%, performance gain is TBD
* For RU > 70%,
  + 1 source [Futurewei] shows performance gain of 73% at CSI feedback overhead A (small overhead)
  + Perormance gain at CSI feedback overhead B (medium overhead) and C (large overhead) is TBD.

For Max Rank 4, performance gain is TBD.

For the evaluation of temporal domain aspects **Case 1** of AI/ML based CSI compression compared to the CSI compression Case 0 *~~benchmark~~ in terms of 5% UPT* *under FTP* traffic, till RAN1 #117,

For Max Rank 1, performance gain is TBD.

For Max Rank 2:

* For RU <= 39%, performance gain is TBD.
* For RU 40-69%, performance gain is TBD
* For RU > 70%,
  + 1 source [Futurewei] shows performance gain of 12% at CSI feedback overhead A (small overhead)
  + Perormance gain at CSI feedback overhead B (medium overhead) and C (large overhead) is TBD.

For Max Rank 4, performance gain is TBD.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Need more time to check results. Would provide comments later. |
| ZTE | The same as last comments. We suggest waiting for more companions’ evaluation results to draw an observation. |

Observation 113a: Full buffer Case 1

Observation 114a: CSI feedback reduction 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 CSI feedback reduction, till RAN1 #117,

* 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 FTP traffic, TBD for full buffer.
  + For CSI feedback overheads B (medium overhead) and C (large overhead), CSI feedback reduction is TBD.
* For Max Rank 4, CSI feedback reduction is TBD.

For the evaluation of temporal domain aspects **Case 1** 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, CSI feedback reduction is TBD.
* For Max Rank 2, CSI feedback reduction is TBD.
* For Max Rank 4, CSI feedback reduction is TBD.

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

### 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*, till RAN1 #117,

For Layer 1,

* 5 sources [Fujitsu, ZTE, Apple, QC, Samsung] observe performance gain of 11-21% at CSI payload X (small payload)
* 1 source [ZTE] observes performance gain of 13.2% at CSI payload Y (medium payload)
* 1 source [ZTE] observes performance gain of 8.9% at CSI payload Z (large payload)

For Layer 2,

* 2 sources [QC, Samsung] observe performance gain between 18-33% at CSI payload X (small payload)
* Performance gains at CSI payload Y (medium payload) areTBD
* Performance gains at CSI payload Z (large payload) are TBD

For Layer 3,

* 1 source [Samsung] observes performance gain between 29-39% 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 Layer 4,

* 1 source [Samsung] observes performance gain between 38-40% 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 2** of AI/ML based CSI compression compared to the CSI compression Case 0 *~~benchmark~~ in terms of SGCS*, till RAN1 #117,

For Layer 1,

* 6 sources [Fujitsu, ZTE, Apple, QC, ViVo, Samsung] observe performance gain of 1-15% at CSI payload X (small payload)
* 1 source [ZTE] observes performance gain of 4.5% at CSI payload Y (medium payload)
* 1 source [ZTE] observes performance gain of 1.1% at CSI payload Z (large payload)

For Layer 2,

* 2 sources [QC, Samsung] observe performance gain of 1-6.7% 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 Layer 3,

* 1 source [Samsung] observes performance gain between 21-30% 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 Layer 4,

* 1 source [Samsung] observes performance gain between 8-9% 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

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, 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.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Need more time to check results. Would provide comments later. |
| Huawei, HiSilicon | Please add Huawei results to the SGCS observation. Our SGCS results are submitted to the FTP/full buffer sheets but not repeated to the “Intermediate KPI” sheet. Sorry for the inconvenience.  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*, till RAN1 #117,  For Layer 1,   * 6 ~~5~~ sources [Fujitsu, ZTE, Apple, QC, Samsung, Huawei] observe performance gain of 11-21% at CSI payload X (small payload) * 2 ~~1~~ sources [ZTE, Huawei] observes performance gain of 11%-13.2% at CSI payload Y (medium payload) * 2 ~~1~~ sources [ZTE, Huawei] observes performance gain of 8.9%-9% at CSI payload Z (large payload)   For Layer 2,   * 3 ~~2~~ sources [QC, Samsung, Huawei] observe performance gain between 18-~~33~~ 37% at CSI payload X (small payload) * 1 source [Huawei] observes performance gains at CSI payload Y (medium payload) of 17% ~~areTBD~~ * 1 source [Huawei] observes performance gains at CSI payload Z (large payload) of 18% ~~areTBD~~   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 SGCS*, till RAN1 #117,  For Layer 1,   * 7 ~~6~~ sources [Fujitsu, ZTE, Apple, QC, ViVo, Samsung, Huawei] observe performance gain of 1-15% at CSI payload X (small payload) * 2 ~~1~~ sources [ZTE, Huawei] observe~~s~~ performance gain of 4.5%-5.7% at CSI payload Y (medium payload) * 2 ~~1~~ sources [ZTE, Huawei] observe~~s~~ performance gain of 1.1%-4% at CSI payload Z (large payload)   For Layer 2,   * 3 ~~2~~ sources [QC, Samsung, Huawei] observe performance gain of 1-~~6.7~~ 20% at CSI payload X (small payload) * 1 source [Huawei] observes performance gain at CSI payload Y (medium payload) of 11.4% ~~is TBD~~ * 1 source [Huawei] observes performance gain at CSI payload Z (large payload) of 8.4% ~~is TBD~~ |
| Samsung | Thank you FL. Please notice that we have considered a different input type.  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. 1 source (Samsung) considered precoding matrix information in angle-delay (W2) domain with spartial and frequency-domain basis vectors reported per N>1 reports. * 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 | We suggest collecting more evaluation results from companies and draw an observation especially the case few sources provides the results. |
| OPPO | Thanks FL. It is better if modify the performance gain range from 11%-21% to 10%-21%, and add our results as follows:  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*, till RAN1 #117,  For Layer 1,   * 6 sources [Fujitsu, ZTE, Apple, QC, Samsung, OPPO] observe performance gain of 10-21% at CSI payload X (small payload)   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 SGCS*, till RAN1 #117,  For Layer 1,   * 7 sources [Fujitsu, ZTE, Apple, QC, ViVo, Samsung, OPPO] observe performance gain of 1-15% at CSI payload X (small payload) |
| CATT | Thanks FL for the summary. Please add CATT results to the SGCS observation. We have updated our SGCS results in “Intermediate KPI” sheet.  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*, till RAN1 #117,  For Layer 1,   * 5 sources [Fujitsu, ZTE, Apple, QC, Samsung] observe performance gain of 11-21% at CSI payload X (small payload) * 1 source [ZTE,CATT] observes performance gain of 13.2%-35% at CSI payload Y (medium payload) * 1 source [ZTE,CATT] observes performance gain of 8.9%-22% at CSI payload Z (large payload)   For Layer 2,   * 2 sources [QC, Samsung] observe performance gain between 18-33% at CSI payload X (small payload) * 1 sources [CATT] observe performance gain at 69% at CSI payload Y (medium payload) ~~areTBD~~ * 1 sources [CATT] observe performance gain at 44% at CSI payload Z (large payload) ~~are TBD~~   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 SGCS*, till RAN1 #117,  For Layer 1,   * 6 sources [Fujitsu, ZTE, Apple, QC, ViVo, Samsung] observe performance gain of 1-15% at CSI payload X (small payload) * ~~1~~2 source [ZTE,CATT] observes performance gain of 4.5%-21% at CSI payload Y (medium payload) * ~~1~~2 source [ZTE,CATT] observes performance gain of 1.1%-17% at CSI payload Z (large payload)   For Layer 2,   * 2 sources [QC, Samsung] observe performance gain of 1-6.7% at CSI payload X (small payload) * 1 sources [CATT] observe performance gain at 48% at CSI payload Y (medium payload) ~~areTBD~~ * 1 sources [CATT] observe performance gain at 41% at CSI payload Z (large payload) ~~are TBD~~ |

Observation 122a: FTP traffic 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 #117,

For Max Rank 1,

* For RU <= 39%, 1 source [Huawei] observes performance gain of 1-3%:
  + 1 source [Huawei] observes performance gain of 3.4% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes performance gain of 2.4% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes performance gain if 2% at CSI feedback overhead C (large overhead)
* For RU 40-69%, 1 source [Huawei] observed performance gain of 4-6%
  + 1 source [Huawei] observes performance gain of 6% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes performance gain of 4% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes performance gain of 6% at CSI feedback overhead C (large overhead)
* For RU > 70%, 1 source [Huawei] observes performance gain of 8-15%
  + 1 source [Huawei] observed performance gain of 15% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes performance gain of 8% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes performance gain of 8% at CSI feedback overhead C (large overhead)

For Max Rank 2,

* For RU <= 39%, 2 sources [Huawei, Interdigital] observes performance gain of 3-12%:
  + 2 sources [Huawei, Interdigital] observe performance gain of 9.6-12% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes performance gain of 3.3% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes performance gain if 5.6% at CSI feedback overhead C (large overhead)
* For RU 40-69%, 2 sources [Huawei, Interdigital] observed performance gain of 9-17%
  + 2 sources [Huawei, Interdigital] observe performance gain of 17% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes performance gain of 10% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes performance gain of 9% at CSI feedback overhead C (large overhead)
* For RU > 70%, 2 sources [Huawei, Interdigital] observe performance gain of 13-29%
  + 2 [Huawei, Interdigital] sources observed performance gain of 13-29% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes performance gain of 17% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes performance gain of 17% at CSI feedback overhead C (large overhead)
* For Max Rank 4, performance gain is TBD.

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 #117,

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%, 1 source [Huawei] observed performance gain of 1-3%
  + 1 source [Huawei] observes performance gain of 3% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes performance gain of 1% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes performance gain of 1 at CSI feedback overhead C (large overhead)
* For RU > 70%, 1 source [Huawei] observes performance gain of 1-6%
  + 1 source [Huawei] observed performance gain of 5.6% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes performance gain of 3.3% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes performance gain of 1.3% at CSI feedback overhead C (large overhead)

For Max Rank 2,

* For RU <= 39%, 2 sources [Huawei, Interdigital] observe performance gain of 1-4%:
  + 2 sources[Huawei, Interdigital] observe performance gain of 1-3.3% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes performance gain of 1% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes performance gain if 2% at CSI feedback overhead C (large overhead)
* For RU 40-69%, 2 sources [Huawei, Interdigital] observed performance gain of 1-6.3%
  + 2 sources [Huawei, Interdigital] observe performance gain of 1-6.3% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes performance gain of 4.2% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes performance gain of 2.6% at CSI feedback overhead C (large overhead)
* For RU > 70%, 2 sources [Huawei, Interdigital] observes performance gain of 2-12%
  + 3 sources [Huawei, FutureWei, Interdigital] observed performance gain of 2-12% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes performance gain of 8% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes performance gain of 5% at CSI feedback overhead C (large overhead)

For Max Rank 4, performance gain is TBD.

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 #117,

For Max rank 1, in general the performance gain increases with the increase of RU:

* For RU <= 39%, 1 source [Huawei] observes performance gain of 4-8%:
  + 1 source [Huawei] observes the performance gain of 8% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes the performance gain of 4% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes the performance gain of 4% at CSI feedback overhead C (large overhead)
* For RU between 40-69%, 1 source [Huawei] observes performance gain of 4-12%:
  + 1 source [Huawei] observes the performance gain of 12% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes the performance gain of 8% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes the performance gain of 4% at CSI feedback overhead C (large overhead)
* For RU > 70%, 1 source [Huawei] observes performance gain of 10-28%:
  + 1 source [Huawei] observes the performance gain of 28% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes the performance gain of 12% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes the performance gain of 10% at CSI feedback overhead C (large overhead)

For Max Rank 2, in general the performance gain increases with the increase of RU:

* For RU <= 39%, 2 sources [Huawei, Interdigital] observe performance gain of 8-45%:
  + 2 sources [Huawei, Interdigital] observe the performance gain of 8-45% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes the performance gain of 8% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes the performance gain of 8% at CSI feedback overhead C (large overhead)
* For RU between 40-69%, 2 sources [Huawei, Interdigital] observe performance gain of 9-27%:
  + 2 sources [Huawei, Interdigital] observe the performance gain of 19-27% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes the performance gain of 13% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes the performance gain of 9% at CSI feedback overhead C (large overhead)
* For RU > 70%, 2 sources [Huawei, Interdigital] observe performance gain of 17-73%:
  + 2 sources [Huawei, Interdigital] observe the performance gain of 27-73% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes the performance gain of 23% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes the performance gain of 17% at CSI feedback overhead C (large overhead)

For Max Rank 4, performance gain is TBD.

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 #117,

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%, 1 source [Huawei] observes performance gain of 4-5%:
  + 1 source [Huawei] observes the performance gain of 5% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes the performance gain of 5% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes the performance gain of 4% at CSI feedback overhead C (large overhead)
* For RU > 70%, 1 source [Huawei] observes performance gain of 1-10%:
  + 1 source [Huawei] observes the performance gain of 10% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes the performance gain of 4% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes the performance gain of 1% at CSI feedback overhead C (large overhead)

For Max Rank 2,

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

For Max Rank 4, performance gain is TBD:

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-16 Type II codebook.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Need more time to check results. Would provide comments later. |
| Huawei, HiSilicon | Please see some corrections for Huawei related results:  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 #117,  For Max Rank 1,   * For RU <= 39%, 1 source [Huawei] observes performance gain of 2-3.4% ~~1-3%~~:   + 1 source [Huawei] observes performance gain of 3.4% at CSI feedback overhead A (small overhead)   + 1 source [Huawei] observes performance gain of 2.4% at CSI feedback overhead B (medium overhead)   + 1 source [Huawei] observes performance gain if 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 mean UPT* *under FTP* traffic, till RAN1 #117,  ……  For Max Rank 2,   * For RU <= 39%, 2 sources [Huawei, Interdigital] observe performance gain of 1-3.3 ~~4~~%:   + 2 sources[Huawei, Interdigital] observe performance gain of 1-3.3% at CSI feedback overhead A (small overhead)   + 1 source [Huawei] observes performance gain of 1% at CSI feedback overhead B (medium overhead)   + 1 source [Huawei] observes performance gain if 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 non-AI/ML *benchmark in terms of 5% UPT under FTP*, till RAN1 #117,  ……  For Max Rank 2, in general the performance gain increases with the increase of RU:   * For RU <= 39%, 2 sources [Huawei, Interdigital] observe performance gain of 8-45%:   + 2 sources [Huawei, Interdigital] observe the performance gain of 9 ~~8~~-45% at CSI feedback overhead A (small overhead)   + 1 source [Huawei] observes the performance gain of 8% at CSI feedback overhead B (medium overhead)   + 1 source [Huawei] observes the performance gain of 8% at CSI feedback overhead C (large overhead) * For RU > 70%, 3 ~~2~~ sources [Huawei, Interdigital, Futurewei] observe performance gain of 17-73%:   + 3 ~~2~~ sources [Huawei, Interdigital, Futurewei] observe the performance gain of 27-73% at CSI feedback overhead A (small overhead)   + 1 source [Huawei] observes the performance gain of 23% at CSI feedback overhead B (medium overhead)   + 1 source [Huawei] observes the performance gain of 17% at CSI feedback overhead C (large overhead)   …… |
| ZTE | We suggest collecting more evaluation results from companies and draw an observation especially the case few sources provides the results. |

Observation 123a: Full buffer 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 #117,

* For Max Rank 1, 1 source [Huawei] observes performance gains of 18-25%
  + 1 source [Huawei] observes performance gains of 25% at CSI feedback overhead A (small overhead)
  + 1 source [Huawei] observes performance gains of 20% at CSI feedback overhead B (medium overhead)
  + 1 source [Huawei] observes performance gains of 18% at CSI feedback overhead C (large overhead)
* For Max Rank 2, 4 sources [Huawei, Fujitsu, Xiaomi, QC] observe performance gains of 2-30%
  + 4 sources [Huawei, Fujitsu, Xiaomi, QC] observe performance gains of 6-30% at CSI feedback overhead A (small overhead)
  + 2 sources [Huawei, Xiaomi] observe performance gains of 3-23% at CSI feedback overhead B (medium overhead)
  + 2 sources [Huawei, Xiaomi] observe performance gains of 2-24% at CSI feedback overhead C (large overhead)
* For Max Rank 4, performance gains are TBD.

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, 2 sources [Huawei, IIT Kanpur] observe performance gains of 0-13%
  + 1 source [Huawei] observes performance gains of 13% at CSI feedback overhead A (small overhead)
  + 2 sources [Huawei, IIT Kanpur] observe performance gains of 5-12% at CSI feedback overhead B (medium overhead)
  + 2 sources [Huawei, IIT Kanpur] observe performance gains of 0-6% at CSI feedback overhead C (large overhead)
* For Max Rank 2, 4 sources [Huawei, Fujitsu, Xiaomi, QC] observe performance gains of 1-14%
  + 4 sources [Huawei, Fujitsu, Xiaomi, QC] observe performance gains of 2-14% at CSI feedback overhead A (small overhead)
  + 2 sources [Huawei, Xiaomi] observe performance gains of 1-14% at CSI feedback overhead B (medium overhead)
  + 2 sources [Huawei, Xiaomi] observe performance gains of 1-9% at CSI feedback overhead C (large overhead)
* For Max Rank 4, performance gains are TBD.

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 #117,

* For Max Rank 1, the performance gains are TBD.
* For Max Rank 2, 2 sources [Fujitsu, QC] observes performance gains of 11-58%
  + 2 sources [Fujitsu, QC] observe performance gains of 11-58% at CSI feedback overhead A (small overhead)
  + Performance gains at medium/large overheads are TBD.
* For Max Rank 4, performance gains are TBD.

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 #117,

* For Max Rank 1, the performance gains are 0-14%.
  + Performance gains at CSI feedback overhead A (small overhead) are TBD
  + 1 sources [IIT Kanpur] observe performance gains of 12-14% at CSI feedback overhead B (medium overhead)
  + 1 sources [IIT Kanpur] observe performance gains of 0-5% at CSI feedback overhead C (large overhead)
* For Max Rank 2, 2 sources [QC, Fujitsu] observes performance gains of 2-35%
  + 2 sources [QC, Fujitsu] observe performance gains of 2.4-35% at CSI feedback overhead A (small overhead)
  + Performance gains at medium/large overheads are TBD.
* For Max Rank 4, performance gains are TBD.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Need more time to check results. Would provide comments later. |
| Huawei, HiSilicon | Please see minor correction for Huawei related results:  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, 2 sources [Huawei, IIT Kanpur] observe performance gains of 0-13%   + 1 source [Huawei] observes performance gains of 13% at CSI feedback overhead A (small overhead)   + 2 sources [Huawei, IIT Kanpur] observe performance gains of 5-12% at CSI feedback overhead B (medium overhead)   + 2 sources [Huawei, IIT Kanpur] observe performance gains of 0-6.6% at CSI feedback overhead C (large overhead)   …… |
| ZTE | We suggest collecting more evaluation results from companies and draw an observation especially the case few sources provides the results. |

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 #117,

- For Max rank = 1,

- For CSI feedback overhead A (small overhead), CSI feedback reduction is TBD;

- For CSI feedback overhead B (medium overhead), CSI-feedback reduction is TBD;

- For CSI feedback overhead C (large overhead), 1 sources [Huawei] observe the CSI feedback reduction of 73-80% for FTP traffic, and 1 sources observes the CSI feedback reduction of 75% for full buffer;

- For Max rank = 2,

- For CSI feedback overhead A (small overhead), CSI feedback reduction is TBD;

- For CSI feedback overhead B (medium overhead), 1 source [QC] observes CSI-feedback reduction of 80% for full buffer;

- For CSI feedback overhead C (large overhead), 1 sources [Huawei] observe the CSI feedback reduction of 53-70% for FTP traffic, and 2 sources [Huawei, QC] observes the CSI feedback reduction of 71-80% for full buffer;

- For Max rank = 4, CSI overhead reduction is TBD.

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), CSI feedback reduction is TBD;

- For CSI feedback overhead B (medium overhead), CSI-feedback reduction is TBD;

- For CSI feedback overhead C (large overhead), CSI feedback reduction is TBD for FTP traffic, and 1 source [Huawei] observes the CSI feedback reduction of 55% for full buffer;

- For Max rank = 2,

- For CSI feedback overhead A (small overhead), CSI feedback reduction is TBD;

- For CSI feedback overhead B (medium overhead), 1 source [QC] observes CSI-feedback reduction of 54% for full buffer;

- For CSI feedback overhead C (large overhead), CSI feedback reduction is TBD for FTP traffic, and 2 sources [Huawei, QC] observes the CSI feedback reduction of 53-60% for full buffer;

- For Max rank = 4, CSI overhead reduction is TBD.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Need more time to check results. Would provide comments later. |
| Huawei, HiSilicon | Please see minor correction for Huawei related results:  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 #117,  - For Max rank = 1,  - For CSI feedback overhead A (small overhead), CSI feedback reduction is TBD;  - For CSI feedback overhead B (medium overhead), CSI-feedback reduction is TBD;  - For CSI feedback overhead C (large overhead), 1 sources [Huawei] observe the CSI feedback reduction of 73-80% for FTP traffic, and 1 sources [Huawei] observes the CSI feedback reduction of 75% for full buffer; |
| ZTE | We suggest collecting more evaluation results from companies and draw an observation especially the case few sources provides the results. |

### 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*, till RAN1 #117,

* 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 *~~benchmark~~ in terms of SGCS*, till RAN1 #117,

* For Layer 1,
  + 1 source [Fujitsu] observes performance gain of 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 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/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* |
| IIT Kanpur | Thankyou moderator for capturing the results. The above assumption bullet needs a correction:   * Benchmark is ~~Rel-16 Type II codebook~~ CSI compression Case 0.   Also, we would like to mention that we have assumed Spatial Consistency Procedure-A in our dataset generation. Hence, if this point can be captured in the assumptions, it will give more clarity. This is applicable to our results of Case2 as well.  We see that many companies have not assumed spatial consistency, and in our observation, spatial consistency has a big impact on the temporal domain compression results. |
| OPPO | Thanks FL. For this item, we observe 1.7% SGCS gain compared to Case 0. So can we modify this item to:  For the evaluation of temporal domain aspects **Case 5** of AI/ML based CSI compression compared to the CSI compression Case 0 *~~benchmark~~ in terms of SGCS*, till RAN1 #117,   * For Layer 1,   + 2 sources [Fujitsu, OPPO] observes performance gain of 1.7-6.3% at CSI payload X (small payload) |

Observation 152a: FTP Case 5

Observation 153a: Full buffer 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 mean UPT under full buffer*, till RAN1 #117,

* For Max Rank 1, performance gains are TBD.
* For Max Rank 2, performance gains are TBD.
* For Max Rank 4, performance gains are TBD.

For the evaluation of temporal domain aspects Case 5 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, 1 source [IIT Kanpur] observes performance gains of 1-6%.
  + Performance gain at CSI feedback overhead A (small overhead) is TBD.
  + 1 source [IIT Kanpur] observes performance gain of 5.69% at CSI feedback overhead B (medium overhead)
  + 1 source [IIT Kanpur] observes performance gain of 1.1% at CSI feedback overhead C (large overhead)
* For Max Rank 2, performance gains are TBD.
* For Max Rank 4, performance gains are TBD.

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 5% UPT under full buffer*, till RAN1 #117,

* For Max Rank 1, performance gains are TBD.
* For Max Rank 2, performance gains are TBD.
* For Max Rank 4, performance gains are TBD.

For the evaluation of temporal domain aspects **Case 5** of AI/ML based CSI compression compared to the CSI compression Case 0 *~~benchmark~~, in terms of 5% UPT under full buffer*, till RAN1 #117,

* For Max Rank 1, 1 source [IIT Kanpur] observes performance gains of 4-14%.
  + Performance gain at CSI feedback overhead A (small overhead) is TBD.
  + 1 source [IIT Kanpur] observes performance gain of 14.08% at CSI feedback overhead B (medium overhead)
  + 1 source [IIT Kanpur] observes performance gain of 4.54% at CSI feedback overhead C (large overhead)
* For Max Rank 2, performance gains are TBD.
* For Max Rank 4, performance gains are TBD.

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

Observation 154a: CSI feedback reduction 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 CSI feedback reduction, till RAN1 #117,

* For Max Rank 1, CSI feedback reduction is TBD.
* For Max Rank 2, CSI feedback reduction is TBD.
* For Max Rank 4, CSI feedback reduction is TBD.

For the evaluation of temporal domain aspects **Case 5** 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, CSI feedback reduction is TBD.
* For Max Rank 2, CSI feedback reduction is TBD.
* For Max Rank 4, CSI feedback reduction is TBD.

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 CSI feedback overhead reduction for Max rank 1/2/4.

- Benchmark is Rel-16 Type II codebook.

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

### 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*, till RAN1 #117, for the mixed scenario of 80% indoor and 20% outdoor users:

For Layer 1,

- 4 sources [oppo, vivo, QC, Fujitsu] observe the performance gain of 1-20% at CSI payload X (small payload);

- 1 source [CATT] observes the performance gain of 0.1% at CSI payload Y (medium payload);

- 1 source [CATT] observes the performance gain of -3.5% at CSI payload Z (large payload).

For Layer 2,

- 1 source [QC] observes the performance gain of 16% at CSI payload X (small payload);

- 1 source [CATT] observes the performance gain of 5.1% at CSI payload Y (medium payload);

- 1 source [CATT] observes the performance gain of -1.1% 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*, till RAN1 #117, for the scenario of 100% outdoor users:

For Layer 1,

- 3 sources [Fujitsu, ZTE, Ericsson] observe the performance gain of 5-12% at CSI payload X (small payload);

- 3 sources [Samsung, ZTE, Ericsson] observe the performance gain of 5-11% at CSI payload Y (medium payload);

- 2 sources [Samsung, ZTE] observe the performance gain of 8-10% at CSI payload Z (large payload).

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

Observation 132a: FTP traffic 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 #117, Performance gains are TBD.

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*, till RAN1 #117, performance gains are TBD.

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

Observation 133a: Full buffer 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 #117, for the mixed scenario of 80% indoor and 20% outdoor users,

- For Max rank 1, the performance gain is TBD.

- For Max rank 2, 1 source [QC] observes the performance gain of 9.4%

- 1 source [QC] observes the performance gain of 9.4% at CSI feedback overhead A (small overhead);

- The performance gain at CSI feedback overhead B (medium overhead) is TBD;

- The performance gain at CSI feedback overhead C (large overhead) is TBD;

- For Max rank 4, the performance gain is TBD.

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 #117, for the mixed scenario of 80% indoor and 20% outdoor users,

- For Max rank 1, the performance gain is TBD.

- For Max rank 2, 1 source [QC] observes the performance gain of 6.9%

- 1 source [QC] observes the performance gain of 6.9% at CSI feedback overhead A (small overhead);

- The performance gain at CSI feedback overhead B (medium overhead) is TBD;

- The performance gain at CSI feedback overhead C (large overhead) is TBD;

- For Max rank 4, the performance gain is TBD.

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* |
|  |  |
|  |  |

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 #117, for the mixed scenario of 80% indoor and 20% outdoor users,

- For Max rank 1, the performance gain is TBD.

- For Max rank 2,

- For CSI feedback overhead A (small overhead), 1 source [QC] observes the CSI feedback reduction of 68%;

- For CSI feedback overhead B (medium overhead), CSI feedback reduction is TBD;

- For CSI feedback overhead C (large overhead), CSI feedback reduction is TBD;

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

### 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 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* |
|  |  |
|  |  |

Observation 142a: FTP traffic Case 4

For the evaluation of temporal domain aspects **Case 4** of AI/ML based CSI compression compared to the *benchmark in terms of mean UPT* *under FTP* traffic, till RAN1 #117, Performance gains are TBD.

For the evaluation of temporal domain aspects **Case 4** of AI/ML based CSI compression compared to the *benchmark in terms of 5% UPT under FTP*, till RAN1 #117, performance gains are TBD.

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

Observation 143a: Full buffer Case 4

For the evaluation of temporal domain aspects **Case 4** of AI/ML based CSI compression compared to the *benchmark, in terms of mean UPT under full buffer*, till RAN1 #117, performance gains are TBD.

For the evaluation of temporal domain aspects **Case 4** of AI/ML based CSI compression compared to the *benchmark in terms of 5% UPT under full buffer*, till RAN1 #117, performance gains are TBD.

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

Observation 144a: CSI feedback reduction Case 4

For the evaluation of temporal domain aspects **Case 4** of AI/ML based CSI compression, compared to the benchmark, in terms of CSI feedback reduction, till RAN1 #117, performance gains are TBD.

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

### 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.

ViVo

1. **Towards the evaluation of localized model, consider the model trained on dataset without spatial consistency as the “generalized model”.**
2. **Localized models should be evaluated on different local regions and the average gain over legacy eType-II can be calculated on these local regions. Results show that stable average performance gain can be observed as long as the number of local regions is large (e.g., 21 cells).**

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.**

Proposal 8: Regarding the data collection for C\SI 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.**

## Discussions

### Local model metric reporting

Proposal 11a:

In case companies model N different local regions and train N different localized models for each region, companies may report the average performance over the N local regions. Companies to report the value of N.

|  |  |
| --- | --- |
| *Support / Can accept* | Lenovo, Futurewei, Huawei, HiSilicon (comment), Panasonic, ETRI, ZTE, Fujitsu, LG |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Support |
| Huawei, HiSilicon | One follow up question to the UPT results: whether/how do we need avoid the randomness by setting a large number of N for localized model? In legacy simulation, 21 sectors are considered. For localized model, we may also need to align the value of N (e.g., 3, 7) rather than being reported by companies. |
| ZTE | Fine with the proposal. |

Proposal 12a:

When reporting the results for the global model, for both the non-AI/ML benchmark and the global AI/ML model, test should be performed on the same test dataset(s) used for the local model(s).

|  |  |
| --- | --- |
| *Support / Can accept* | Lenovo, Huawei, HiSilicon (comment), Panasonic, ZTE, LG |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | The baseline for global model is the case without spatial consistency, as discussed in the next proposal. Otherwise, it is difficult to construct a huge dataset that can be used for global model training. |
| Huawei, HiSilicon | OK for the AI/ML performance. But one more question for the non-AI/ML benchmark performance: we noticed that some company consider the common SGCS value of benchmark (eType II) for both global model and local model, while some company consider different SGCS values of benchmark between global model and local model, and also different SGCS values of benchmark over different local cases.  How should we set the benchmark? In our understanding, the fluctuation/randomness of benchmark performance should be avoided, since non-AI/ML solution is not data driven. |
| ETRI | We agree on the direction. Our question for clarification is that whether the test dataset for the global model is the superset of all the test dataset of the local models? In that case, the word ‘same’ need to be modified, in our view. |
| ZTE | Fine with the proposal. |

### Local region modeling

Proposal 13a:

For the evaluation of AI/ML-based CSI compression using localized models in Release 19, additionally consider the following option as a starting point to model global and local regions:

* Option 3: 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.
  + The dataset for global region is generated by disabling spatial consistency.

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.

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Lenovo | In doing so, proposal 12a is not possible anymore since the test data for the global and local model would be different. |
| vivo | Support this proposal. Maybe this is a further elaboration of Option1? |
| Huawei, HiSilicon | The only difference from Option 1 is “The dataset for global region is generated by disabling spatial consistency”. In our understanding, even enabling spatial consistency, we can also achieve global dataset with generalized performance, so it seems no need to mandate disabling spatial consistency. |
| Panasonic | We would like to clarify the motivation to consider Option 3 in addition to Option 1. |
| ETRI | Our question is that this proposal is for obtaining only training dataset for the global model, or obtaining both training and test dataset? |
| ZTE | To our understanding, the sub-bullet is for Option 1, not Option 3.   * Option ~~3~~1: The dataset is derived from UEs dropped within the local region, with spatial consistency modelling as per TR 38.901.   In addition, we are not clear the motivation of the proposal whether Option 1 has priority over Option 2? From our perspective, both options can achieve the assumptions for localized model and no priority issue should be adopted for both cases. |
| LG | The clarification is needed whether spatial consistency is enabled or not for the dataset for global region. To my understanding, the spatial consistency is already considered for global region and then, is it for the purpose on performance comparision? |
| OPPO | Not very clear why we need option 3? |

### Results collection for temporal aspects with localized models

Proposal 14a:

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)** |

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Support |
|  |  |

### Evaluation results

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 #117

For Layer 1,

* + 2 sources [ZTE, ViVo] observe the performance gain of 22-28% over benchmark, at CSI payload X (small payload), using TSF compression
  + 1 source [ZTE] observes the performance gain of 16 % at CSI payload Y (medium payload), using TSF compression
  + 1 source [ZTE] observes the performance gain of 9 % at CSI payload Z (large payload), using TSF compression

For Layer 2,

* + 1 source [ZTE] observes a performance gain of 34% over benchmark, at CSI payload of X (small payload)
  + 1 source [ZTE] observes a performance gain of 19% over benchmark, at CSI payload of Y (medium payload)
  + 1 source [ZTE] observes a performance gain of 10% 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 #117,

For Layer 1,

* + 2 source observes the performance gain of 1-22% [ZTE ViVo] over global model, at CSI payload X (small payload), using TSF compression
  + 1 source [ZTE] observes the performance gain of 12 % at CSI payload Y (medium payload), using TSF compression
  + 1 source [ZTE] observes the performance gain of 2 % at CSI payload Z (large payload), using TSF compression

For Layer 2,

* + 1 source [ZTE] observes a performance gain of 27% over global model, at CSI payload of X (small payload)
  + 1 source [ZTE] observes a performance gain of 12% over global model, at CSI payload of Y (medium payload)
  + 1 source [ZTE] observes a performance gain of 0.11% over global model, at CSI payload of Z (large payload)

Note: No companies submitted evaluation results with the localized models having lower complexity than the global models, till RAN1 #117.

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* |
| vivo | Need more time to check results. Would provide comments later. |
| ZTE | Companies needs to report how to train and testing dataset generation, e.g., considering time factors. |
| Intel | Same comment as for temporal CSI compression:  In the current release we are looking into the tradeoff between perfoamance and complexity/overhead. So, we think that range of AI/ML model complexities shall be captured in observations (applies to all the observation with performance gain and complexity saving). |

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 #117

For Layer 1,

* + 1 source [ViVo] observes a performance gain of 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 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 #117,

For Layer 1,

* + 2 sources [ViVo, Nokia] observes a performance gain of 0-5% 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 0-3% over global model, at CSI payload of Y (medium payload)

Note: No companies submitted evaluation results with the localized models having lower complexity than the global models, till RAN1 #117.

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* |
|  |  |
|  |  |

### 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

Table 2-1: Comparison among various options for inter-vendor collaboration issue

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Collaboration complexity** | **Performance** | **Interoperability and testing aspects** | **Feasibility** |
| **Option 1**  (Fully standardize reference model) | Least complexity among all the options | May have worse performance in the field compared to other options. | Least testing complexity involved among all the options. | Least feasible among all the options from specification effort perspective. |
| **Option 2**  (Standardize dataset) | Deprioritized | | | |
| **Option 3a**  (Standardized reference model structure + Parameter exchange between NW-side and UE-side; parameters received at the UE or UE-side goes through offline engineering at the UE-side) | * Less complexity if parameter exchange method and procedure are standardized. * More complexity and effort otherwise. | * May have better performance than Option 1. * May have performance limitation compared to Option 4 and Option 5. | * May be feasible for RAN4 testing assuming vendor can develop the model based on the specified structure. * Can be better determined by RAN4. | * Less time in agreeing reference model structure only compared to Option 1. * Significant effort needed in discussing the method(s) and procedures for model parameter exchanges and alignment of quantization. |
| **Option 3b**  (Standardized reference model structure + Parameter exchange between NW-side and UE-side; parameters received at the UE are directly used for inference at the UE without offline engineering) | * Less complexity if parameter exchange method and procedure are standardized. * More complexity and effort otherwise. | May have worse performance than Option 3a. | May be more feasible from testing perspective than Option 3a. | * Same as Option 3a, less time in agreeing reference model structure only compared to Option 1. * Significant effort in discussing the procedures for model parameter exchanges, i.e., Case z4. |
| **Option 4**  (Standardized data / dataset format + Dataset exchange between NW-side and UE-side) | * Less complexity if dataset exchange method and procedure are standardized. * More complexity and effort otherwise. | May have better performance than Option 1. | * May be feasible for RAN4 testing assuming vendor can develop the model based on the specified dataset format and the dataset received. * Can be better determined by RAN4. | * Less time in agreeing data or dataset format only compared to Option 1, Option 2, and Option 3. * Significant effort needed in discussing dataset format and content and the mechanism for quantization alignment. |
| **Option 5a**  (Standardized model format + Reference model exchange between NW-side and UE-side; model received at the UE or UE-side goes through offline engineering at the UE-side) | * Less complexity if model exchange method and procedure are standardized. * More complexity and effort otherwise. | May have better performance than Option 1, Option 2, and Option 3. | * May be feasible for RAN4 testing assuming vendor can develop the model based on the specified model format. * Can be better determined by RAN4. | * Less time in agreeing model format only compared to Option 1, Option 2, and Option 3. * Significant effort needed in discussing the method(s) and procedures for model exchange and alignment of quantization. |
| **Option 5b**  (Standardized model format + Reference model exchange between NW-side and UE-side; model received at the UE are directly used for inference at the UE without offline engineering) | * Less complexity if model exchange method and procedure are standardized. * More complexity and effort otherwise. | May have worse performance than Option 5a. | May be more feasible from testing perspective than Option 5a. | * Same as Option 5a, less time in agreeing model format only compared to compared to Option 1, Option 2, and Option 3. * Significant effort in discussing the procedures for model exchange, i.e., Case z4 and in negotiation between vendors regarding model structure alignment. |

***Proposal 1: Among the options to alleviate/resolve the issues related to inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model, consider further studying Option 3, Option 4, and Option 5.***

***FFS: data/dataset format, content, and model format.***

***FFS: mechanisms/procedures for model parameters/model and dataset exchange.***

Tejas Networks Ltd.

**Proposal 3: 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 4:** **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 5: 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 6: 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 7: 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 8: 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.**

Ericsson

1. For RAN1 Option 1, 3, 4, 5, study performance monitoring mechanisms including the feasibility and complexity, for ensuring the performance of two-sided models when operating in the field.
2. For RAN1 Option 1, 3, 4, 5, study how to detect root cause of faulty performance for CSI compression using two-sided models.
3. RAN1 study methods to further improve two-sided model performance based on standardized reference model (i.e., RAN1 Option 1 and RAN 4 Option 3/4) and still maintain interoperability, including at least the following aspects:

* UE-side and/or NW-side independently improve their actual UE-part model and NW-part model using a dataset B dataset, which is different from the dataset A used for defining the standardized reference model.
* Monitoring mechanisms, including feasibility and complexity, to monitor the performance of the enhanced two-side model when operating in the field.
* How can the operator identify the responsibility if the enhanced two-sided model fails in operation in the field?

1. For inter-vendor training collaboration, deprioritize 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.
2. Deprioritize Option 5a for inter-vendor training collaboration.
3. Deprioritize option 3a-2 for parameters exchanged from the NW-side to UE-side.
4. For Option 3a, deprioritize over-the-air delivery method for exchanging information from the NW-side to UE-side.
5. For Option 3a, further study the following aspects:

* The feasibility and complexity of standardizing the structure of the reference model.
* The necessary information to be exchanged from the NW-side to the UE-side, including at least parameters of the reference model, training/testing dataset or information related to collecting training/testing dataset, and performance target.
* The feasibility and complexity of standardizing at least the format and structure of all information to be exchanged from the NW-side to UE-side.
* The feasibility and complexity of standardizing a delivery method that does not involve over the air delivery.
* The methods, including feasibility and complexity, for supporting a UE to be loaded with different CSI generation models depending on at least which network vendor it connects to, without disclosing NW proprietary information.
* Monitoring mechanisms, including feasibility and complexity, to monitor performance of the enhanced two-side model when operating in the field.
* How can the operator identify the responsibility if the two-sided model fails in operation in the field?
* Comparison among different sub-options (Option 3a-1, 3a-2, and 3a-3).

1. 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.
2. For Option 3b, further study the following aspects:

* 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?
* Monitoring mechanisms, including feasibility and complexity, to monitor performance of the enhanced two-side model when operating in the field.
* How can the operator identify the responsibility if the two-sided model fails in operation in the field?
* The performance comparison between Option 3a and Option 3b.

1. Deprioritize Option 5b for inter-vendor training collaboration.
2. For Option 4, deprioritize over-the-air delivery method for exchanging information from the NW-side to UE-side.
3. Deprioritize Option 4-2 and 4-3 for inter-vendor training collaboration.
4. For Option 4-1, further study the following aspects:

* The necessary information to be exchanged from the NW-side to the UE-side, including at least training/testing dataset and performance target.
* The feasibility and complexity of standardizing at least the format and structure of all information to be exchanged from the NW-side to UE-side.
* The feasibility and complexity of standardizing a delivery method that does not involve over the air delivery.
* The methods, including feasibility and complexity, for supporting a UE to be loaded with different CSI generation models depending on at least which network vendor it connects to, without disclosing NW proprietary information.
* How to check if the trained CSI generation model at the UE-side is compatible with the CSI reconstruction model at the NW-side, and whether the paired model fulfils the performance target？
* Monitoring mechanisms, including feasibility and complexity, to monitor performance of the enhanced two-side model when operating in the field.
* How can the operator identify the responsibility if the two-sided model fails in operation in the field?

Huawei

***Proposal 7: Capture Table 9 to the TR 38.843 for the comparison over the model exchange related options/sub-options of inter-vendor training collaboration.***

Table 9 Comparison over model exchange related options/sub-options

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model exchange related options/sub-options** | | | **Inter-vendor collaboration complexity** | **Performance** | **Feasibility** | **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 8: Capture Table 11 to the TR 38.843 for the comparison over the dataset exchange related sub-options of inter-vendor training collaboration.***

Table 11 Comparison over dataset exchange related method sub-options

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset exchange related sub-options** | | **Inter-vendor collaboration complexity** | **Performance** | **Feasibility** | **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 |

Intel

***Proposal 3***:

* *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.*

***Proposal 4***:

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

***Proposal 5***:

* *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 6***:

* *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 7***:

* *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 8***:

* *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.*

|  |  |  |
| --- | --- | --- |
|  | **For UE-part training** | **For NW-part training** |
| **Option 1** | CSI generation part is specified:  **Type 1 NW side training**.  CSI reconstruction part is specified:  **Type 2 sequential training**. | CSI generation part is specified:  **Type 2 sequential training**.  CSI reconstruction part is specified:  **Type 1 UE side training**. |
| **Option 2** | **Type 3 NW-first** | **Type 3 UE-first** |
| **Option 3a-1** | CSI generation part weights are transferred/delivered from NW side to UE side:  **Type 1 NW side training**. | |
| **Option 3a-2** | CSI reconstruction part weights are transferred/delivered from NW side to UE side:  **Type 3 starting with NW side training**. | |
| **Option 3a-3** | CSI generation part weights are transferred/delivered from NW side to UE side:  **Type 1 NW side training**. | |
| **Option 3b** | CSI generation part weights are transferred/delivered from NW side to UE side:  **Type 1 NW side training**. | |
| **Option 4-1/4-2/4-3** | Dataset is transferred/delivered from NW side to UE side: **Type 3 NW-first**. | |
| **Option 5a-1** | CSI generation part model is transferred/delivered from NW side to UE side:  **Type 1 NW side training**. | |
| **Option 5a-2** | CSI generation part model is transferred/delivered from NW side to UE side:  **Type 3 starting with NW side training**. | |
| **Option 5a-3** | CSI generation part model is transferred/delivered from NW side to UE side:  **Type 1 NW side training**. | |
| **Option 5b** | CSI generation part model is transferred/delivered from NW side to UE side:  **Type 1 NW side training**. | |

***Proposal 9***:

* *Training collaboration types are further discussed based on progress achieved in the Rel-18 SI.*
  + *Re-discuss entries in Table 5.1-1 and Table 5.1-2 of TR 38.843 where consensus has not been reached.*

***Proposal 10***:

* *For CSI compression using two-sided model use case, the following changes (in red) are endorsed for the table with the pros/cons of training collaboration type 1.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Training type  Characteristics | Type1: NW side | | Type 1: UE side | |
| Unknown model structure at UE | Known model structure at UE | Unknown model structure at NW | Known model structure at NW |
| Extendibility: to train new UE-side model compatible with NW-side model in use; | Yes | Yes | Yes | Yes |
| Extendibility: To train new NW-side model compatible with UE-side model in use | Yes | Yes | Yes | Yes |

***Proposal 11***:

* *For CSI compression using two-sided model use case, the following changes (in red) are endorsed for the table with the pros/cons of training collaboration types 2 and 3.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Training types  Characteristics | Type 2 | | Type 3 | |
| Simultaneous | Sequential  NW first (note 1) | NW first | UE first |
| Feasibility of allowing UE side and NW side to develop/update models separately | Infeasible | Feasible | Feasible | Feasible |
| Extendibility: to train new UE-side model compatible with NW-side model in use; | Not support | Support | Support | Support |
| Extendibility: To train new NW-side model compatible with UE-side model in use | Not support | Not Support | Support | Support |

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.***

Spreadtrum

***Observation 2: For options 3 and option 5,*** ***although there are performance gains compared to option1, it also introduces the complexity of model/parameter exchange.***

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

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

Samsung

***Proposal#11: 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#12: Deprioritize Option 5b as***

* **It does not alleviate the inter-vendor collaboration complexity due to the requirement on offline inter-vendor collaboration for model structure alignment**
* **UE and/or network may be required to support multiple model structures**

***Proposal#13: 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#14: 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.***

vivo

1. **For option 1, RAN1 considers specifying CSI generation part.**
2. **Companies supporting option 5b should clarify whether such case is included in option 5b: UE indicates NW the supported model structure via a standardized procedure.**
3. **For the study of inter-vendor collaboration issue, consider the timescale of deployment when comparing different options.**
4. **RAN1 concludes that it is recommended to support option 3b to address inter-vendor training collaboration for CSI compression**
5. **RAN1 considers specifying the reference model structure as a starting point for the specification of a reference model.**
6. **Towards the specification of reference model structure, 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.**

CATT

**Proposal 6: For Option 1, to avoid duplicate work between RAN1 and RAN4, study the difference aspects of RAN1 Option 1 and RAN4 option 3/RAN 4 option 4 by RAN1 (e.g., the model at which sides should be standardized), and not to study the feasible and methodology(s) on fully standardized reference model in RAN1.**

**Proposal 7: 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 8: 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 9: To alleviate / resolve the issues related to 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 10: To alleviate / resolve the issues related to inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model, prioritize the solutions with over-the-air signaling standardized.

Proposal 11: To alleviate / resolve the issues related to 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  Related to RAN4-Option3 and RAN4-Option4 | Feasible for UE implementation  Large spec effort (initially & for evolution) |
| **Option 2** | Low / None | Limited | Good  Related to RAN4-Option4 | Feasible for UE implementation.  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 | Limited  Worse than Options 1/2 if the UE-side would retrain/re-deploy a model different from the one transferred from the NW-side to the UE-side, and dataset/data collection information is not transferred from the NW-side to UE-side, otherwise, better than Options 1/2. | Poor interoperability (need RAN4 study) | 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 imply the model  Large spec effort (initially & for evolution) but less than Options 1/2 |
| **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 | Limited  Better than Options 1/ 2 | Poor interoperability (need RAN4 study) | Feasible for UEs capable of updating parameters directly  Large spec effort (initially & for evolution) but less than Options 1/2 |
| **Option 4** | Low/none if dataset is exchanged through the over-the-air signaling  Medium/high if dataset is exchanged offline | Best  Better than Options 1/2/3 | Poor interoperability (need RAN4 study) | Feasible for UE implementation  FFS the impacts of the delay from the time the UE/UE-side receives the parameters to the time the UE can imply 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 | Limited (worse than Options 1/2/4) if the UE-side would retrain/re-deploy a model different from the one transferred from the NW-side to the UE-side, and dataset/data collection information is not transferred from the NW-side to UE-side.  Otherwise, Best. | Poor interoperability (need RAN4 study) | 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 imply the model  Less spec effort than Options 1/2/3 |
| **Option 5b** | Medium/high | Best | Poor interoperability (need RAN4 study) | Feasible for UEs capable of updating parameters directly  Less spec effort than Options 1/2/3 |

China Telecom

***Proposal 2: Prioritize Option 3 for inter-vendor training collaboration for further study.***

***Proposal 3: For option 4 and option 5, definition of data/dataset and model format should be discussed first.***

CMCC

*Observation 3: Option 5a need more time for model deployment compared to Option 5b.*

*Observation 4: The performance of Option 5a may be better than Option 3b.*

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

*Observation 6: The multi-vendor collaboration issue still exists for Option 5a and 5b.*

*Observation 7: Both Option 5a and Option 5b refer to model transfer/delivery Case z4.*

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 direct 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.**

Lenovo

1. Due to performance limitation and also required high specification effort, we suggest deprioritizing Option 2 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. Therefore, 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. Prioritize schemes based on exchange of complete model (or options based on dataset exchange) over options based on exchange of model parameters only.
8. Capture the following table comparing different options based on different criteria. The most important negative and positive points are highlighted with red and green color, respectively.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | Required signalling overhead | Required specification efforts | Possibility of Performance Check | Offline engineering |
| Encoder model | With offline engineering | 3a-1 | Low | High | No | Needed |
| 5a-1 | Moderate | Low | No | Needed |
| 4-1 | High | Moderate | No | Needed |
| No offline engineering | 3b | Low | High | No | No need |
| 5b | Moderate | Low | No | No need |
| Decoder model | With offline engineering | 3a-2 | Low | High | Yes | Needed |
| **5a-2** | Moderate | Low | Yes | Needed |
| **4-2** | High | Moderate | Yes | Needed |
| Both encoder and decoder | With offline engineering | 3a-3 | Sum of overheads needed for respective ()-1 and ()-2 cases | Sum of the specification efforts needed for respective ()-1 and ()-2 cases | Yes | Needed |
| 5a-3 | Yes | Needed |
| 4-3 | Yes | Needed |

1. 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).

Nvidia

**Proposal 5: Study further Options 1, 3, 4, and 5 for inter-vendor training collaboration.**

LGE

**Proposal #6: Prioritize Option 4 for addressing inter-vendor training collaboration.**

**Proposal #7. Study on model complexity method, e.g., knowledge distillation, to further reduce the CSI training/signaling complexity for Type 3 training collaboration.**

SKT

1. For AI/ML-based CSI compression study using two-sided model in Release 19, Inter-vendor training collaboration option 4 and 5 are prioritized than others.

Fujistu

***Proposal 3:***

* *The parameters/dataset/model should be transferred via air interface to reduce inter-vendor collaboration complexity for Options 3/4/5, respectively.*

***Proposal 4:***

* *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 5:***

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

***Proposal 6:***

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

***Proposal 7:***

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

***Proposal 8:***

* *To alleviate the inter-vendor training collaboration issues, Option 4 is preferred compared with Option 3 and Option 5.*

***Proposal 9:***

* *RAN 1 to further study the data format in Option 4.*

Xiaomi

***Proposal 1: For Option 3a/5a/4, the parameter/model//dataset and additional information are directly exchanged from NW-side to OTT server by standardizing signalling between NW-side to OTT server to save transmission resource and UE’s power.***

NEC

***Proposal 5: Capture the following characteristics of Option 1~5:***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Option type  Characteristics | Option 1 | Option 2 | Option 3 | Option 4 | Option 5 |
| Inter-vendor collaboration complexity | None | Little | More than Option 1/2. | More than Option 1/2. | More than Option 1/2. |
| Performance | Limited. | Limited. | Not limited. | Not limited.  Upper limit is better than Option 3. | Not limited.  Upper limit is better than Option 3. |
| Interoperability and RAN4 / testing related aspects | Align with Option 3/4 in RAN4 | \ | Align with Option 3/4 in RAN4 | \ | Align with Option 3/4 in RAN4 |
| Feasibility | Heavy standardization works. | Heavy standardization works. | Few standardization works.  Online delivery of parameter is needed. | Few standardization works.  Online dataset delivery is needed. | Few standardization works.  Online model delivery/transfer is needed. |

***Proposal 6:*** ***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 7: For Option 3a/5a and Option 3b/5b:***

* ***Time required to deploy a model that can be used for model inference: Option 3b/5b is less than Option 3a/5a***
* ***Performance: Option 3a/5a is better than Option 3b/5b***

***Proposal 8: 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.***
* ***Option 3a-3/5a-3: Model/Parameters exchanged from the NW-side to UE-side are both CSI generation part and CSI reconstruction part.***

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***

ZTE

***Proposal 7:*** *For Option 1, prioritize to study the case of standardized CSI generation part.*

***Proposal 8:*** *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.*

***Proposal 9:*** *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 10:*** *For Option 3/4/5, prioritize to study over-the-air delivery scheme.*

***Proposal 11:*** *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.*

CAICT

***Proposal 1: The overall analysis on alleviate/resolve issues related to inter-vendor training collaboration of options3/4/5 are listed in the table.***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Inter-vendor collaboration complexity** | **Performance** | **Interoperability** | **Feasibility** |
| **Option 3a** | **High** | **Good** | **Specified model structure and parameters should be considered** | **Feasible with potential high overhead** |
| **Option 3b** | **Medium** | **Good (Pre-deployment test should be considered)** | **Specified model structure and parameters should be considered** | **Feasible** |
| **Option 4** | **High** | **Good** | **Standardized dataset is required to ensure interoperability** | **FFS** |
| **Option 5a** | **High** | **Good** | **Specified reference model should be considered** | **Feasible with potential high overhead** |
| **Option 5b** | **Medium** | **Good (Pre-deployment test should be considered)** | **Specified reference model should be considered** | **Feasible** |

Panasonic

**Proposal 1: 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. | 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. | 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) | 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 | 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) | 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. | 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. | May be feasible, but more standardization effort than Option 1 may be needed. |

OPPO

*Proposal 10: 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 11: RAN1 further study how to standardize reference model structure in Option 3.*

*Proposal 12: RAN1 further study how to standardize data / dataset format in Option 4.*

*Proposal 13: RAN1 further study how to standardize reference model format in Option 5.*

Nokia

Proposal 6: RAN1 needs to quickly perform cost-benefit analysis on Option 5 compared to other contending options, i.e., Option 3 and Option 4, and to consider deprioritization of Option 5 in case there is no clear advantage with respect to the required standardization/inter-vendor operability support efforts.

Proposal 7: RAN1 and RAN4 should align on reference CSI reconstruction structure, wherever applicable, to streamline specification of the reference decoder for UE conformance testing (RAN4) or for deployment in the field (RAN1), in case the respective working group determines to specify the reference model structure for its own purpose.

Proposal 8: As regards interoperability and inter-vendor collaboration aspects investigation, RAN1 can focus on Option 4 first, and possibly extend its specification effort to Option 3, if the need arises.

Proposal 9: As regards Option 4, RAN1 should consider schemes to align quantization rule between UE-side and NW-side, in view of support of inter-vendor collaboration. RAN1 should also consider specification of quantization rule and associated VQ codebook, if applicable, in case out-of-3GPP alignment of vendor-specific proprietary quantization rule could not bring about significant performance gain.

Proposal 10: RAN1 should consider allowing 3GPP members an access to training dataset/reference model parameters/API, etc., irrespective of their original contributors for the benefit of the whole eco system.

Proposal 11: For the inter-vendor training cooperation in CSI compression, RAN1 shall study additional metric or information (besides mere data pair of original CSI and codeword) to monitor and guide UE-side encoder’s model quality in NW-first sequential separate training framework, and eventually boost CSI compression performance and minimize inter-vendor collaboration complexity.

Docomo

**Proposal 1**

* **If RAN4 specifies the reference model, the RAN1 Option 1 can be supported with little specification effort.**
* **Study the enhancements under the scope of RAN1 Option 1 for performance enhancements, e.g., define RAN1 enhancements based on the reference model.**

**Proposal 2**

* **Deprioritize the study of using Option 4 as a stand-alone inter-vendor collaboration scheme.**

MTK

1. If downselection is needed, prioritize UE-first sequential separate training.
2. Discuss how option 1 and option 3 will ease inter-operability test efforts and also candidate scenarios/config as starting point of tests.
3. Discuss bounding performance deviation of AI/ML models derived from fully standardized AI/ML to ensure inter-vendor compatibility
4. If area-specific AI/ML models show promising performance, prioritize option 3 to make are-specific models feasible.

Qualcomm Incorporated

***Proposal 1: Conclude that proprietary models and proprietary exchange methods are feasible. Inter-vendor collaboration complexity can be addressed by standardization of model design aspects or leveraging any standardization (e.g., standardized structure) resulting from the listed inter-vendor collaboration options.***

***Proposal 2: For parameter / model exchange of CSI reconstruction part with performance requirement to facilitate UE-side offline engineering of developing CSI generation part, Conclude that inter-vendor collaboration complexity and interoperability can be addressed with following consideration on specification impact (either or both)***

* Specify over-the-air signalling for decoder transfer.
* Offline exchange methods with standardized procedure (if needed) and / or standardized model design aspects (for option 5a) such as quantization, payload, layer/rank-specific models, scalability to various subband / port configurations, etc.

***Proposal 3: For parameter / model exchange of CSI reconstruction part, conclude that***

* UE side may further go through offline engineering of developing CSI generation part or run inference on device directly
* inter-vendor collaboration complexity can be addressed via (either or both)
  + Specified over-the-air signalling for parameter / model transfer.
  + Offline exchange methods with standardized procedure (if needed) and / or standardized model design aspects (for option 5a) such as quantization, payload, layer/rank-specific models, scalability to various subband / port configurations, etc
* Feasibility issue and performance limitation due to UE distribution mismatch can be addressed by inter-vendor collaboration.

***Proposal 4: Consider unified framework and signalling for CSI generation part and CSI reconstruction part exchange with and without UE side offline engineering.***

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:**

1. **Model architecture**
2. **Number of layers**
3. **Hyper-parameters**
4. **Quantization method**

## Discussions

|  |  |
| --- | --- |
| Futurewei | Further study option 3 / 4 / 5 |
| Ericsson | Generally, for RAN1 option 1 / 3 / 4 / 5,   * study performance monitoring mechanism to ensure two-sided model performance in the field. * Study how to detect root cause of performance failure. * Study UE / NW side independent improve performance on dataset B which is different from dataset A which is used to train the standardized reference model.   Additionally,   * For 3a and 4-1,   + study feasibility and complexity of standardization of model structure,   + necessary information exchanged,   + Signaling not involve OTA,   + model switching based on NW vendor without disclosing NW proprietary information, * For 3b,   + study feasibility and complexity of standardization of model structure,   + standardize model parameter precision and pre-processing.   Deprioritize 3a-2, 5a, 5b, 4-2, 4-3. |
| Huawei | Capture comparison table among all the options and sub-options, with following highlights.   * Inter-vendor collaboration can be relieved in 3a/3b/4 using standardized procedure, but not for 5a/5b due to model structure alignment. * 3a performance is more limited than 3b. * 4-2 performance is limited (feedback bits, reconstructed CSI) |
| Intel | For 3a/5a, study offline delivery for case y and OTA signaling for case z4.  Option 3b is preferred over 3a from inter-vendor collaboration perspective.  Performance target provided to UE may not need to be achieved, instead, study it as threshold/target for LCM.  Map options with training types in Rel-18. |
| TCL | Further down-selection among option 3, 4, 5. Option 3 and 5 preferred if model is exchanged from NW to UE. |
| Samsung | Deprioritize option 5b.  For 4-1/2/3, further study whether target CSI is shared before or after quantization. |
| vivo | * For option 1, specify CSI generation part. * Consider timescale of deployment when comparing different options * RAN conclude support 3b to address inter-vendor collaboration complexity. * For 3b, * Pre-deployment testing can be done. * Necessity of post deployment testing can be studied. * Easily support localized model * For 3a, * Ota signaling complexity less than offline manner. * Offline signaling may involve more vendor. * Difficult to do pre-deployment testing. * Long time scale for deployment * Propose procedures for specification of reference model structure. |
| CATT | Study difference between ran1 option1 vs. ran4 option 3 / 4.  Prioritize encoder parameter transfer in option 3.  In option 4, prioritize 4-1 (target CSI and feedback bits)  Deprioritize offline signaling between servers, prioritize OTA signaling. |
| China Telecom | Prioritize option 3.  Clarify dataset / model format in option 4 and 5 |
| Sony | Support option 3b, 3a can be considered as additional retraining of encoder |
| Lenovo | Prioritize 3a-2 / 5a-2 / 4-2 (decoder or decoder data sharing), because allows UE to evaluate the performance of the trained encoder.  Deprioritize 3a-3, 4-3, 5a-3 as they will not outperform 3a-2, 5-2 and 4-2.  3b may have issue of yielding unacceptable performance, prioritize 3a over 3b. |
| LGE | Prioritize option 4 |
| SKT | Prioritize option 4 and 5 |
| Fujistu | OTA signaling for option 3 / 4 / 5 to reduce inter-vendor collaboration complexity.  Prioritize 3a-2 / 5a-2 over 3a-1, 5a-1, 3a-3, 5a-3  Prefer option 4 over option 3 / 5. Among the latter, prefer 3a/5a over 3b/5b. |
| Xiaomi | In option 3a/5a/4, directly exchange between servers. |
| NEC | 3a has larger timescale than 3b for deployment, but with better performance.  Prioritize 3a-2/5a-2, 3a-3/5a-3 with decoder part model / parameter exchange |
| Google | Prioritize option 3 and 5, whether offline engineering or direct inference is upto UE implementation |
| ZTE | For option 1, standardize encoder.  For option 3 / 4 / 5, prioritize OTA signaling.  For option 3 / 4 / 5, further study post-development operation to guarantee interoperability and performance |
| OPPO | Distinguish the reference model in RAN1 and RAN4: RAN1 has higher requirement, ran1 cannot directly use the reference model in RAN1 |
| Nokia | * RAN1 should quickly evaluate whether deprioritize option 5 as it has no clear benefit than others. * RAN1 and RAN4 to align the purpose the specified model so as to determine which working group to specify the model. * RAN1 can potentially leverage RAN4 work for the model structure in option 3. * RAN1 should consider alignment on quantization rule including VQ codebook, incase proprietary rule could not bring about a good gain. * Consider allowing central registry for access to dataset/model/API. * Consider additional metric to guide UE offline training, e.g., MSE-SGCS relation (MSE is on the latent) |
| DCM | Study enhancement based on the reference model.  Deprioritize option 4. |
| Qualcomm | * Proprietary methods can also alleviate inter-vendor collaboration complexity with specified model design aspects. * Offline engineering with decoder sharing is feasible with OTA signaling or offline signaling considering model design aspects aligned. * Offline engineering with encoder sharing is also possible with OTA signaling or offline signaling, feasibility and performance limitation can be resolved with additional inter-vendor collaboration (e.g., multi-encoder training w/ UE vendor involved, testing on target UEs) * Unified framework and signaling for encoder / decoder sharing |
| CEWiT | For option 1/3, the model structure includes, model architecture, #layers, hyper-parameters, quantization method.  Support model ID based method for configuration, identification, parameter transfer. |

Signaling to exchange parameter / model / dataset

Unified signaling

Deprioritization

Post-deployment testing / monitoring

Distinguish RAN1 option 1 with RAN4 option 3

Aspects of specifying model structure

|  |  |  |
| --- | --- | --- |
|  | Yes | no |
| 3a-1  (10 vs. 1) | E: spec effort, offline signaling, monitoring, model switching, generalization  HW: relieved with OTA signaling, limited perf than 3b  Intel: offline delivery, OTA  Vivo: OTA, hard for pre-testing, long time scale deployment  CATT: OTA  CT  Sony  Xiaomi: offline server-server  Google: 3a or 3b is upto UE implementation  QC: OTA, offline, unified framework for enc/dec sharing | Lenovo |
| 3a-2  (11 vs. 1) | E: spec effort, offline signaling, monitoring, model switching, generalization  HW: relieved with OTA signaling, limited perf than 3b  Intel: offline delivery, OTA  Vivo: OTA, hard for pre-testing, long time scale deployment  CT  Lenovo: preferred over 3b  Fujistu: 3a/5a-2 > 3a/5a-1/3 > 3b/5b, OTA signaling  Xiaomi: offline server-server  NEC  Google: a or b is upto UE implementation  QC: OTA, offline, unified framework for enc/dec sharing | E: not support fallback using transferred encoder |
| 3a-3  (10 vs. 1) | E: spec effort, offline signaling, monitoring, model switching, generalization  HW: relieved with OTA signaling, limited perf than 3b  Intel: offline delivery, OTA  Vivo: OTA, hard for pre-testing, long time scale deployment  CT  Sony  Xiaomi: offline server-server  NEC  Google: a or b is upto UE implementation  QC: OTA, offline, unified framework for enc/dec sharing | Lenovo |
| 3b  (8) | E: spec effort, parameter precision, pre-processing  HW: relieved with OTA signaling  Intel: offline delivery, OTA, preferred over 3a  Vivo: address inter-vendor collaboration  Sony  NEC  Google: a or b is upto UE implementation  QC: OTA, need further inter-vendor collaboration for enc design and testing |  |
| 4-1  (6 vs. 2) | E: spec effort, offline signaling, monitoring, model switching, generalization  CATT  LGE  SKT  Fujistu: better than option 3 / 5, OTA signaling  Xiaomi: offline server-server | Lenovo,  DCM |
| 4-2  (5 vs. 3) | Lenovo  LGE  SKT  Fujistu: better than option 3 / 5, OTA signaling  Xiaomi: offline server-server | E  Huawei: performance limited  DCM |
| 4-3  (4 vs. 3) | LGE  SKT  Fujistu: better than option 3 / 5, OTA signaling  Xiaomi: offline server-server | E,  Lenovo,  DCM |
| 5a-1  (5 vs. 4) | Intel: offline delivery, OTA  SKT  Xiaomi: offline server-server  Google: a or b is upto UE implementation  QC: OTA, offline, unified framework for enc/dec sharing | E,  Lenovo,  Huawei (structure alignment),  Nokia |
| 5a-2  (7 vs. 3) | Intel: offline delivery, OTA  Lenovo  SKT  Fujistu: 3a/5a-2 > 3a/5a-1/3 > 3b/5b, OTA signaling  Xiaomi: offline server-server  Google: a or b is upto UE implementation  QC: OTA, offline, unified framework for enc/dec sharing | E,  Huawei (structure alignment),  Nokia |
| 5a-3  (5 vs. 4) | Intel: offline delivery, OTA  SKT  Xiaomi: offline server-server  Google: a or b is upto UE implementation  QC: OTA, offline, unified framework for enc/dec sharing | E,  Lenovo,  Huawei (structure alignment),  Nokia |
| 5b  (2 vs. 3) | Sony  SKT  Google: a or b is upto UE implementation | E,  SS,  Nokia |

As summarized in the above table, option 3 receives higher attention than option 4 and 5. Companies elaborated their views on whether CSI generation part or CSI reconstruction part or both should be exchanged, provided details on what signalling can be used to achieve the parameter transfer, and also discussed the expected performance in the field.

From FL perspective, the number of proponents of CSI generation part or CSI reconstruction part exchange are comparable. The main argument of exchanging CSI generation part or parameters is that at least UE can directly implement it in a short time scale. In the meanwhile, UE can also process offline-engineering to optimize or re-develop the CSI generation part according to their own capability and dataset. On the other hand, some companies mention that exchanging the CSI reconstruction part would allow UE to evaluate the performance of the optimized CSI generation part. Besides, exchanging CSI reconstruction part parameters would allow UE-side to develop a better CSI generation part based on UE-specific data distribution.

From signalling aspects, some companies think standardized over-the-air signaling can be used to address inter-vendor collaboration complexity in 3a/3b/5a/5b, while some others think offline signaling is proper for 3a/5a/4 because the communication is between NW side and UE side servers.

From these perspective, for progress, FL thinks it would be helpful to converge to a common ground among the options. We may want to conclude positively how inter-vendor collaboration can be addressed, while the details of sub-options including specification impacts can be decided further under the common framework.

### Over-the-air exchange for Option 3/4/5

Proposal 21a:

Conclude that **over-the-air signalling** can be used for parameter / model exchange in option 3a/5a and 3b to address the inter-vendor collaboration complexity. Study the necessity and feasibility of following potential specification impact

* The transferred parameter / model may be CSI generation part, CSI reconstruction part or both, with potentially common signalling or framework
* Assistant information, e.g., performance requirements, etc
* Timescale for the model successfully deployed at UE
* Model structure for which the parameters are exchanged in option 3a/3b, and the associated pre-processing and parameter precision

|  |  |
| --- | --- |
| *Support / Can accept* | IIT Kanpur, |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Lenovo | Okay with 3a/3b, but 5a might need more clrafication on why it needs to be over the air. |
| vivo | Support. |
| Futurewei | In general, we are ok with the direction but we think we need agreement from companies regarding the first sentence of the proposal before discussing potential specification impact; thus, we suggest discussin the first part in this proposal:  Conclude that **over-the-air signalling** can be used for parameter / model exchange in option 3a/5a and 3b to address the inter-vendor collaboration complexity. |
| Huawei, HiSilicon | 1), we think the first sentence and the remaining part are stating two aspects. First sentence is describing the inter-vendor collaboration, while the rest part is describing the feasibility aspect. Therefore, this proposal can be  2), for the 1st sentence, the key message should be that over the air signalling can alleviate inter-vendor collaboration, while offline signalling can hardly alleviate inter-vendor collaboration, regardless of the options. Therefore, we do not need  Proposal 21a-1:  Conclude that **over-the-air signalling** ~~can be used~~ for parameter / model / dataset exchange ~~in option 3a/5a and 3b~~ ~~to~~ can address the inter-vendor collaboration complexity.  3), for the second part, which focuses on model/parameter exchange, we may start with delivering the CSI generation part. As we are discussing the model transfer from NW to UE, the CSI reconstruction part should be NW proprietary – this guarantees the NW vendors are motivated to develop competitive decoder to achieve high end to end performance.  4), for Option 3a/3b, we should sort out what are the aspects to converge standardized model structure.  5), for Option 5a, we should not forget the dataset for the UE side to retrain/update the UE part model. Without the training dataset, the trained UE part model may still fail to pair with the NW part model, since the dataset used for training NW part model is different from dataset used for retraining/updating the UE part model.  6), BTW, Option 3a seems unnecessary to standardize model structure since anyway UE side will retrain the model.  Proposal 21a-2:  Study the necessity and feasibility of following potential specification impact   * The transferred parameter / model may be CSI generation part as a starting pint~~, CSI reconstruction part or both, with potentially common signalling or framework~~ * Assistant information, e.g., performance requirements, Dataset etc, delivered for retraining/updating in Option 5a * Timescale for the model successfully deployed at UE * Model structure for which the parameters are exchanged in option ~~3a/~~3b, and the associated pre-processing and parameter precision. Model structure related study includes at least:   + - Scalability: Scalability method, e.g., adaptation layer, truncation/zero padding, etc.     - Quantization method: Format of quantization, e.g., scalar quantization, vector quantization.     - Model backbone: CNN, MLP, Transformer, LSTM + Transformer, etc.   Layer related parameters: Number of layers, depth of layer, activation function, etc. |
| Panasonic | We support Futurewei’s suggestion. |
| Samsung | What is the difference between 3b (OTA by definition) and 3a OTA? |
| ETRI | We need to disscuss on the first part at first.  In our view, OTA signalings for exchanging model/parameters for options that require additional offline engineering (i.e., 3a/5a) are not appropriate due to the model development timeline may much longer than the case w/o the additional offline engineering. |
| ZTE | To our understanding, the main bullet is also applicable to Option 5b since model exchange is via OTA signalling. So, we suggest adding Option 5b in the main bullet. To be aligned with the wording in WID, “address” needs to be modified to “resolve/alleviate”.  For the 1st bullet, we can understand the intention, however, **we need to first discuss and have a common understanding on the applicable scenarios for exchanging CSI generation part / CSI reconstruction part / both parts before discussing proposal.** To our understanding, to specify which part of model structure has one-to-one correspondence to the sub-options of Option 3a.We need to discuss these issues based on the basic common understanding achieved by companies.  Additionally, for the 2nd sub-bullet, “Assistant information” and “performance requirements” are not aligned with the previous agreement, we suggest changing it into “Additional information” and “Performance target”.  For the 4th sub-bullet, we are not clear about “the associated pre-processing and parameter precision”, which needs further clarification.  The modified version is listed as  Conclude that **over-the-air signalling** can be used for parameter / model exchange in option 3a/5a and 3b/5b to ~~address~~alleviate/resolve the inter-vendor collaboration complexity. Study the necessity and feasibility of following potential specification impact   * The transferred parameter / model may be CSI generation part, CSI reconstruction part or both, with potentially common signalling or framework * ~~Assistant~~Additional information, e.g., performance ~~requirements~~ targets, etc * Timescale for the model successfully deployed at UE   Model structure for which the parameters are exchanged in option 3a/3b, and the associated pre-processing and parameter precision |
| Fujitsu | We think Option 4 should also be included in the proposal since over-the-air exchange is also possible for Option 4. |
| SK telecom | We think that over-the-air signalling can be used for option 4 (data set) also. Is there any reason to exclude option 4 in this proposal? If not, we suggest adding Option 4 as follow:  Conclude that **over-the-air signalling** can be used for parameter / data set / model exchange in option 3a/5a, 4 and 3b to address the inter-vendor collaboration complexity. |
| Ericsson | For option 3a/5a, we don’t think over-the-air signaling is feasible, necessary to address the intervendor collaboration complexity. Option 3a/5a are about parameter/model exchange from NW side to UE side, to assist the UE-side offline engineering for model training and performance testing, the NW side also needs to provide performance target and training/testing dataset or information related to collecting training/testing dataset to the UE-side. The feasibility, necesserty and complexity of using over-the-air for exchanging these information has not been properly studied and justified.  While over-the-air signaling may be used for option 3b to address the inter-vendor training collaboration complexity, there are other potential issues (e.g., UE implementation feasibility) that requires further study for this option. Hence, we suggest to have separate discussions for option 3a/5a and 5b.  In addition, inter-vendor collaboration complexity is not only caused via model training but also performance monitoring and faulty detection (e.g., encoder vs. decodeor vs. training/monitoring data mismatch issue).  Without proper study of the above issues, we cannot conclude that over-the-air signaling can address inter-vendor collaboration complexity. |
| Sony | We support the direction of proposal.  However, we also support Futurewei’s suggestion. |
| Intel | In our view the wording ‘Conclude that…’ seems too strong. It appears like it is feasible already while not aspects have been considered so far to conclude that.  We propose the following wording change:  Conclude that **over-the-air signalling** can be used in principle for parameter / model exchange in option 3a/5a and 3b to address the inter-vendor collaboration complexity |
| OPPO | We prefer to first consider the main bullet:  Conclude that **over-the-air signalling** can be used for parameter / model exchange in option 3a/5a and 3b to address the inter-vendor collaboration complexity.  The necessity and feasibility issue can be discussed separately. |
| CATT | We are generally ok with the FL’s direction. Similar to Futurewei, we think it is better to separate the conclusion and another proposal for the OTA signalling further study spec impact. |

### Offline exchange for Option 3/4/5

Proposal 22a:

Conclude that **offline signalling** with standardized procedure / signalling or standardized model design aspects can be used for parameter / model / dataset exchange in option 3a/5a and 4 to address the inter-vendor collaboration complexity.

* Further study the potential standardized procedure or signalling, e.g., transferred content, assistant information, etc
* Further study standardized model design aspects, e.g., quantization rule, layer/rank common/specific model, etc

|  |  |
| --- | --- |
| *Support / Can accept* | Lenovo (comment) |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Lenovo | We are generally okay with the proposal with the following clarfications on offline signaling.  Conclude that **offline ~~signalling~~ exchange of information** with standardized procedure / signalling or standardized model design aspects **with possibly over the air indications** can be used for parameter / model / dataset exchange in option 3a/5a and 4 to address the inter-vendor collaboration complexity.  The intension is that, in some cases the model/parameter/dataset might be stored offline at a node, but accessing the correct model/parameter/dataset may need over-the-air signalling. So, not all steps are offline; only the model/parameter/dataset exchange is offline. |
| vivo | Need more discussion on this proposal.  Offine signaling may still have vendor to vendor specific issues which incurs inter-vendor collaboration complexity. The conclusion might be on the other direction. |
| Futurewei | Our understanding regarding “offline signalling” is that this is for offline collaboration between vendors; thus, we suggest modifying the wording:  Conclude that **offline ~~signalling~~** collaboration with standardized procedure / signalling or standardized model design aspects can be used for parameter / model / dataset exchange in option 3a/5a and 4 to address the inter-vendor collaboration complexity. |
| Huawei, HiSilicon | “Offline signalling” is contradictory with “standardized procedure/signaling” – if the two vendors offline exchange the dataset/model, then they do not need standardized procedure/signalling, while on the other hand the inter-vendor collaboration is not allevieated.  We think before discussing the path/signalling of delivery (which is more related with RAN2), we may first focus on the content of the dataset/model delivery. After we understand the rough size of model/ dataset, and the information that needs to be delivered, we then will have better understanding on the signalling.  Therefore, we suggest the following formulation:  ~~Conclude that~~ **~~offline signalling~~** ~~with standardized procedure / signalling or standardized model design aspects can be used for parameter / model / dataset exchange in option 3a/5a and 4 to address the inter-vendor collaboration complexity.~~   * Further study the potential standardized procedure or signalling for dataset delivery of Option 4, e.g., transferred content, assistant information, etc, including at least:   + Data format: 1) The type of ground-truth CSI (eigenvectors, channel matrix, angular-delay domain eigenvectors, channel matrix, etc.), the dimension of input (port number, bandwidth, etc.)/output (CSI payload size). 2) The type of the CSI feedback.   + Dataset construction: 1) Number of data samples in the dataset. 2) Dataset split/segmentation   + Scalability: The construction of data samples for a set of scalability configuration   + Quantization method: Format of quantization, e.g., scalar quantization, vector quantization * Further study standardized model design aspects, ~~e.g., quantization rule, layer/rank common/specific model, etc~~ including at least:   + Quantization of the parameters   + layer/rank common/specific model   + Scalability: Scalability method, e.g., adaptation layer, truncation/zero padding, etc.   + Quantization method: Format of quantization, e.g., scalar quantization, vector quantization.   + Model backbone: CNN, MLP, Transformer, LSTM + Transformer, etc.   + Layer related parameters: Number of layers, depth of layer, activation function, etc. |
| Panasonic | We think it would be necessary to clarify that the “offline signaling” in the proposal does not intend offline collaboration between vendors. |
| Samsung | Same view as other companies above. |
| ZTE | Agree with the above companies. |
| Fujitsu | The offline singling option needs more discussion. As other companies mentioned, the inter-vendor collaboration issue still exists for such option. |
| LG | Clarification on ‘offline signaling’ is needed to expedite the further discussion |
| Ericsson | Inter vendor collaboration complexity is not only caused via model training but also performance monitoring and faulty detection (e.g., encoder vs. decodeor vs. training/monitoring data mismatch issue). Hence, without porpoer study on the performance monitoring and faulty detection aspects for these options, we cannot conclude that offline signalling with standardized procedure / signalling or standardized model design aspects can be used to address the inter-vendor collaboration complexity. The first sentence shall be removed, more studied on performance monitoring and fault detection should be done for these options. |
| Sony | We support Panasonic’s suggestion.  The “offline signaling” should be clarified. |
| CATT | Share the same view as other companies. Further clarifications are needed. The proposal is a bit confusing regarding offline signalling with standardized procedure / signalling or standardized model design aspects. |

### Option 5

Regarding option 5, the main difference compared to option 3 is how to align model structure. Some companies think aligning model structure offline would not completely alleviate inter-vendor collaboration. Nokia proposes that RAN1 should make quick analysis and conclude whether to continue pursuing along this route. Thus, from FL perspective, for progress, we may want to list the issues of option 5 for further study, and also conclude that option 5 can be achieved in a proprietary manner. While standardization of model structure / parameter and its exchange is studied in Rel-19, it should be noted that the use of proprietary model and proprietary model/parameter exchange is also possible and should be allowed if vendors want to collaborate.

Proposal 23a:

Conclude that the use of proprietary model structures and proprietary model/parameter exchange is possible.

The relevant standardization of LCM (e.g., model identification, model design aspects, inference and monitoring) should not preclude proprietary models and proprietary exchanges.

In light of the use of proprietary model and proprietary model/parameter exchange is also possible, RAN1 to further study the necessity of Option 5.

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Lenovo |  |
| vivo | It is also possible that Option 5 may have another version without propriateary model structure exchange, e.g., through standarzied way of exchanging model structure information. |
| Huawei, HiSilicon | We understand the motivation of delivery proprietary model, but the question is: whether/how it is possible to alleviate the inter-vendor collaboration?  The two vendors need to anyway align the model representation format (MRF), e.g., pytorch, tensorflow, onnx, etc. There is no precedent that 3gpp incorporates an open source community, and it requires tremendous efforts to specify a MRF at 3GPP/RAN1. So, realistically, the two sides probably need to offline align the MRF. |
| ZTE | For the 1st and 3rd bullet, we think there is no clear evidence to conclude the use is possible. If we need to progress Option 5, we need further discuss the feasibility of the use of proprietary model/model structures and proprietary model/parameter exchange. So, we suggest the revision as  ~~Conclude~~Further study the feasibility that the use of proprietary model structures and proprietary model/parameter exchange ~~is possible~~.  The relevant standardization of LCM (e.g., model identification, model design aspects, inference and monitoring) should not preclude proprietary models and proprietary exchanges.  In light of the use of proprietary model and proprietary model/parameter exchange ~~is also possible~~, RAN1 to further study the feasibility and necessity of Option 5. |
| Fujitsu | Generally fine with the proposal. |
| Ericsson | For Option 5a, a NW vendor needs to offline align with multiple UE/chipset vendors at least on the definition of reference model (i.e., 5a-1, 5a-2, or 5a-3) and the reference model format, and to make sure that the selected model format (if not developed in 3GPP) is suitable in terms of supported functionalities. In addition, each UE vendor is required to collaborate with different NW vendors to develop and test multiple CSI generation models to be compatible with different CSI reconstruction models from different NW vendors. All these alignment implying high multi-vendor collaboration complexity.  For Option 5b, a NW vendor and a UE vendor need to offline align the model structure of the CSI generation model, which implies high complexity of inter-vendor bilateral collaboration, similar to Option 5a. As the model structured is offline aligned between NW and UE vendors, it is likely that the model parameter design including parameter precision also requires offline vendor-vendor specific offline alignment to make it feasible for UE implementation.  As option 5a/b requires significant offline multi-vendor collaboration for model training and performance testing, this option should not be part of the 3GPP discussion at all.  Hence, we suggest to deprioritize option 5a/b in 3GPP discussions |
| CATT | We propose the following revision for the 1st sentence:  Conclude that the use of proprietary model structures and proprietary model/parameter exchange is possible with large inter-vendor collaboration efforts. |

### Option 1

Regarding option 1, while we clarified that it corresponds to RAN4 option 3 or 4 in last meeting, some companies want to clarify how to proceed in RAN1 by differentiating with RAN4. In light of previous meeting agreement and considering companies proposal, following proposal may be considered.

Proposal 24a:

For Option 1 (corresponding to RAN4 option 3/4), conclude that the standardized reference model in RAN4 and its compatible counterpart model can be used in the field as a baseline scheme or a fallback scheme.

* Note: the actual model used at UE or NW-side may be different from the standardized reference model.
* Note: The structure of the standardized reference model from Option 1 (RAN4 option 3/4) may be used for Option 3.

Further study field performance of Option 1 by considering that UE-side and/or NW-side independently improve their actual UE-part model and NW-part model using a dataset B (representing field data), which is different from the dataset A (representing dataset for developing the standardized reference model).

|  |  |
| --- | --- |
| *Support / Can accept* | Samsung |
| *Object / Have a concern* | Huawei, HiSilicon |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Lenovo | The word baseline is not very clear here for us. Does the following wording is inline with the FL intension.  For Option 1 (corresponding to RAN4 option 3/4), conclude that any model used in filed should be matched with the standardized reference model in RAN4 (and its compatible counterpart model) ~~can be used in the field as a baseline scheme or a fallback scheme.~~ |
| vivo | It was concluded last meeting the field performance would degrade for Option1. If the field performance is not justified, we are not sure it should be used as a baseline or fallback scheme for the field.  Moreover, the direction of study to independently optimize the performance from each side currently does not have any results from any company. Stating that we need to put efforts into this direction of study would diverge focus in this already limited time study in Rel-19. |
| Huawei, HiSilicon | Actually, we have concerns that a standardized model at RAN4 for the testing purpose probably cannot work with good performance in field, where the data from practical channel enviroments may be largely different from the testing lab/testing dataset. Thus, the testing can only justify that the UE vendor is capable to develop/train/adopt the UE part model to be paired with the NW part model of a NW/TE vendor, but cannot justify the tested model always works well in field – note, that what RAN4 defines is minimum requirement, which means the standardized model may come from one of the lowest performances from all companies. That is to say, the tested model may not serve as a baseline/fallback scheme (at least higher than eType II) in field.  Same reason, that the RAN4 determined model structure cannot be adopted for Option 3 due to its limited performance. There are a number of companies in RAN4 proposing MLP/CNN. For RAN1, on the other hand, at least for submitted results for localized models, Transformer backbone is adopted. |
| Panasonic | We think the standardized reference model in RAN4 can be used as a starting point for the reference model of Option 1 and some of “minimum performance” are defined by Option 1, Then, we suggest the following change to the main bullet..  For Option 1 (corresponding to RAN4 option 3/4), conclude that the standardized reference model in RAN4 and its compatible counterpart model can be used for defining the minimum performance requirement ~~in the field as a baseline scheme or a fallback scheme~~. |
| Samsung | Agree with FL’s proposal. |
| ZTE | Generally fine with the proposal. We are not clear about the fall-back scheme, which needs to be further clarified. Is it a LCM issue? |
| Fujitsu | We agree with other companies that model specified by RAN4 may not be suitable for RAN1 since RAN4 only defines the minimum performance requirement.  This proposal needs more discussion. |
| Ericsson | Support in general. Suggest to update the first Note as the following:  • Note: the actual model used at UE or NW-side may be different from the standardized reference model, as long as it meet the minimum performance requirement(s) defined in RAN4 or is interoperable with the specified reference model(s). |

### Option 3/4/5 performance under UE data distribution mismatch

For performance comparison, one important aspect for the analysis is that whether the options or sub-options yield a UE vendor agnostic CSI generation part that is trained based on common dataset, or the options or suboptions is able to yield UE / UE vendor specific CSI generation part trained based on UE / UE vendor specific data distribution. Thus, following proposal can be considered to help draw conclusion from performance perspective.

Proposal 25a:

Study performance of option 3 / 4 / 5 and their sub-options under **UE data distribution mismatch** and study solutions to address performance degradation (if any).

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Lenovo | In general it could be applicable to Option 1 as well. |
| vivo | We showed in our previous papers that CSI compression performance is not sensitive to RF mismatch (e.g., Rx antenna spacing, Rx antenna imbalance). On the other hand if such issues need to be studied, similar issues would also need to be considered for NW data distribution mismatch, e.g., when new sites are added.   * + We did some evaluations to see the impact of antenna imbalance and antenna spacing. In Table 2-4, each UE antenna has a different amplitude and phase disturbance, where the power factor of the first antenna is 1 and different random power attenuations are set on other antennas. It could be seen from the performance of different spacings and antenna imbalance is negligible. It is not necessary for the network to maintain different models for different devices.   Table 2-3: The performance of different antenna spacing for CSI compression.   |  |  |  |  | | --- | --- | --- | --- | | Antenna spacing | 0.3 wave length | 0.5 wave length | 0.8 wave length | | Model performance | 0.8721 | 0.8793 | 0.8766 |   Table 2-4: The performance of antenna imbalance for CSI compression.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Power imbalance between antennas | 0.0dB | 0.5dB | 1.0dB | 2.0dB | 3.0dB | | Model performance | 0.8793 | 0.8793 | 0.8791 | 0.8784 | 0.8775 | |
| Futurewei | We share the same view as Lenovo that this proposal is general across all options for inter-vendor collaboration. |
| Huawei, HiSilicon | We are not clear the motivation of this issue. Since we will focus on the NW side train/NW first train, it means the dataset is collected by NW side. If there is a new UE vendor/type joining, the NW side can first collect the new data from the new UE vendor/type, then either to generate a mixed dataset by blendering the new data to the old dataset, or to generate a new dataset specifically for the new UE vendor/type. A new NW part model is then trained based on the mixed dataset or new dataset.  Actually, considering the diversity of data feature caused by UE side antenna is supposed to be much less than the NW side antenna. The UE data mismatch may not be a big problem in real network. |
| Panasonic | We share the same view as Lenovo and Futurewei. |
| ZTE | We suggest companies to study and evaluate whether/how the issue of UE data distribution mismatch would incur performance degradation. |
| Fujitsu | What does it exactly mean by “UE data distribution mismatch”?  Agree with other companies that Option 1 has the same issue. |
| Ericsson | Support in general. And agree with Lenovo and Futurewei that this proposal shall be applied to all options. |
| Sony | We share Lenovo and Futurewei’s view. |
| CATT | Agree with Lenovo |

### Interoperability of Option 3/4/5

One issue raised for option 3 / 4 / 5 (especially for 3a / 4 / 5a) is that the interoperability may not be ensured as UE / UE vendor specific CSI generation part is developed via UE side offline engineering. Hence, according to some companies preference, following proposal is considered.

Proposal 26a:

For option 3 / 4 / 5 and their sub-options, study mechanisms of **post-deployment testing and performance monitoring** to guarantee good performance in the field and potentially identify the cause of performance failure.

|  |  |
| --- | --- |
| *Support / Can accept* | Lenovo, Panasonic |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Support |
| Futurewei | We think whether to study post-deployment testing performance monitoring for all the sub-options under 3/4/5 depends on the discussion outcome of inter-vendor collaboration options first, so our suggestion is to wait till more progress on the options. |
| Huawei, HiSilicon | We agree with the principle that post deployment monitoring is needed, but what is the difference from the generic monitoring procedure? |
| ZTE | Fine with the proposal. |
| Fujitsu | We think this issue could also be applicable to Option 1 and it should be included in the proposal. |
| LG | Fine |
| Ericssion | Fine with the direction.  Based on current RAN1 and RAN4 discussions, there are three possible approaches/outcomes for introduting AI/ML in 3GPP for two-sided models:  Apporach 1: UE encoder model is tested by RAN4/5 and then that model is deployed in the field. Interoperability is achieved. This apporach is aligned with RAN1 Option 1. Basic performance in achieved per RAN4/5 testing definition and procedures. No model updates after deployment is possible.  Apporach 2: No RAN4/5 test of UE encoder models and no IoDT before deployment. No interoperability is achieved but is replaced by monitoring. This apporach is aligned with RAN1 Option 3/4/5. Model updates via information provided from NW-side to UE-side.  Apporach 3: A combination of approach 1 and apporach 2, interoperability is achieved at a basic performance level, localized models or model performance enhancement is achieved by monitoring. This corresponds to using RAN1 option 1 as a starting point but may addtionaly use Option 3/4/5.  We think all approaches/options require various degree of performance monitoring mechanisms to enable fault (encoder vs. decoder. vs. training/inference data mismatch) detection and LCM opterations in the filed. Compared to Apporach1, other Apporaches may require much more complicated monitoring and fault responsibility detection solutions, which is currently lack of proper study in RAN1.  Hence, we suggest update the proposal as:  For option 1/3/4/5, study and compare the required performance monitoring mechnaisms and fault detection solutions, including at least the feasibility and complexity for both UE and NW sides. |
| Sony | We support the proposal. |
| CATT | ok |

### 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.

Huawei

***Proposal 9: 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.***

***Proposal 10: 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 11: For the dataset delivery of CSI compression over air-interface, NW can split 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.***

***Observation 10: The motivation of introducing the assistance information for assisting UE side data categorization is not clear considering the following points:***

* ***UE can train a generalized model that is applicable to multiple scenarios/antenna layouts.***
* ***UE can autonomously identify the scenario without the need for gNB notification.***
* ***The categorization principle and granularity of the scenarios identified by NW side may not match the categorization principle of the UE side.***
  + ***To achieve aligned categorization rule, offline interoperation of the physical meaning (e.g., scenarios, antenna layouts) of the categorization ID between NW side and UE side may be inevitable, which harms the engineering isolation and may probably disclose the proprietary in the end.***

Vivo

1. **RAN1 concludes that it is recommended to use e-type II-like codebook-based quantization with potential extended parameter combinations for NW-side data collection of precoding matrix.**

CATT

Proposal 12: In CSI compression using two-sided model use case, discuss data collection for training as if it would be specified in Rel-19.

Proposal 13: In CSI compression using two-sided model use case, further study the following on data collection for training and data collection for monitoring:

* **Data collection procedure, e.g., UE-side data collection, or NW-side data collection;**
* **Contents of data sample**
  + **Data sample type, e.g., precoding matrix, channel matrix;**
  + **Data sample format: Scalar quantization and/or codebook-based quantization (e.g., Type II alike);**
* **Assistance information, e.g., information for categorizing the data for the purpose of differentiating characteristics of data due to specific configuration, scenarios, site etc;**
  + **Study the necessity and potential solutions (if the necessity has been identified);**
* **Enhancement on CSI-RS configuration** 
  + **Study the necessity and potential solutions (if the necessity has been identified).**

China Telecom

***Proposal 4: Support to enable high-quality data collection from UE to network, at least including:***

* ***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***

CMCC

***Proposal 10: 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 11: 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 12: 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.***

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.

Fujistu

***Proposal 10:***

* *For CSI compression using two-sided models, RAN 1 to further discuss using codebook-like approach to report ground-truth CSI for AI/ML model training, e.g. Rel-16 e-type II-like codebook with enhanced parameter values.*

Xiaomi

***Proposal 2: At least for training Type 1 at NW side and Type 3 with NW-first, it is necessary and feasibility for NW side data collection considering the following aspects:***

* ***Significant feedback overhead reduction by using codebook-based quantization.***
* ***Signalling overhead reduction by using cell-specific CSI-RS resource configuration.***
* ***No much strict latency requirement for data collection for model training or performance monitoring.***

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***

ZTE

***Proposal 13:*** *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 14:*** *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*

Panasonic

**Observation 8: 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 9: 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 10: Assuming fast monitoring is 100s of ms order, U-plane, RRC or MAC-CE can be sufficient.**

**Observation 11: 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 12: 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 13: 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 14: 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.**

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.

Fraunhofer

***Proposal 3: If the NW is collecting the data for training, the UE should provide the NW node with essential information about the configuration of data e.g., type of the CSI included in the data, quantization parameters, how often the data should be collected, etc.***

***Proposal 4: The signaling/configuration for data collection should include a quality requirement/threshold for the ground truth labels. If such requirement is not guaranteed to be met, then data collection should not be initiated.***

## Discussions

|  |  |
| --- | --- |
| Companies | Views |
| Huawei | Confirm necessity for UE to report ground-truth   * Precoder, not channel, high resolution eT2 * Rank indicated by NW   Type 3 training:   * study dataset ID used to differentiate models trained with different dataset, * study dataset size   NW sent different portion of dataset to different UEs, and UE side perform re-combination  Motivation of introducing assistance information for UE side data categorization is unclear. |
| vivo | Study data collection procedure, data content, assistance information, enhancement on CSI-RS |
| China Telecom | UE report data quality, e.g., SINR, CQI  NW configures threshold for data quality and UE reports qualified data |
| CMCC | Enhanced eT2 and R18 doppler CB as starting point |
| Lenovo | Procedure / signaling to associate data samples with conditions / additional conditions  Consider transmitting more samples, rather than less samples with higher resolution especially when overhead is more important |
| Fujistu | Ground-truth reporting using codebook, e.g., enhanced eT2 |
| Xiaomi | Overhead reduction using codebook-based quantization |
| Google | NW indicate the rank  Reporting of singular values  Report CQI/RI in addition to ground-truth  Reuse CPU framework to handle the UE complexity  Support data collection / measured by NW configuration or UE request |
| ZTE | Enhanced eT2 for ground-truth reporting  UE report data quality  NW configure data quality threshold |
| Panasonic | Ground-truth reporting can be via U-plane  High resolution eT2  Time stamps, cell ID, UE location can be considered as UE side additional condition  UE side associated ID is necessary  Configuration ID of NW side additional condition is necessary |
| MTK | SRS can be used for data collection |
| Fraunhofer | NW configure data quality  Content includes data type, quantization parameter, how often data is collected |

From the summary table, it can be seen that following aspects are elaborated and discussed

Format of data sample: enhanced eT2 or R18 CB

Data quality and rank

Association of NW / UE side conditions

Mechanisms of data collection

These aspects were studied in Rel-18, FL thinks the discussion can wait a bit, and in this meeting we can prioritize discussion on aspects that are more crucial toward September checkpoint. Also, companies are encouraged to propose more detailed or concrete proposals compared to Rel-18 TR, so that we can draw constructive conclusions in Rel-19.

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

# Monitoring

## Summary of company proposals

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

Huawei

***Proposal 12: For monitoring metrics, consider intermediate KPI and eventual KPI as the starting point in Rel-19.***

* ***Further discuss the reporting mode, e.g., per sample reporting and statistic reporting over a number of monitored samples.***
* ***Legacy CSI based monitoring and input distribution-based or output distribution-based monitoring can be deprioritized in Rel-19.***

***Proposal 13: There is no strong motivation for specifying the UE side proxy model for monitoring.***

***Proposal 14: For the intermediate KPI based monitoring, consider the signaling of ground-truth CSI/recovery CSI between NW and UE to assist the calculation of the intermediate KPI.***

* ***NW side monitoring based on the ground-truth CSI (target CSI with realistic channel estimation) reported by the UE.***
* ***UE side monitoring based on the recovery CSI (output of the CSI reconstruction model) indicated by NW.***

Intel

***Observation 6***:

* *At least the following aspects require further study for NW-side model performance monitoring based on ground truth CSI quantization:*
  + *Robustness of model performance monitoring against channel variations in time.*
  + *Efficiency of model performance monitoring considering the corresponding CSI feedback overhead.*

***Proposal 12***:

* *Whether/how to support new CSI report format for ground truth CSI quantization should be further studied considering the corresponding CSI overhead.*

***Proposal 13***:

* *NW-side model performance monitoring based on an intermediate KPI calculated using channel measured via SRS can be supported without additional specification impact.*
  + *Target CSI: channel/precoding matrix derived via SRS.*
  + *Output CSI: output of the two-sided model with channel/precoding matrix derived via SRS at the input.*

***Proposal 14***:

* *For SRS-based model performance monitoring, reuse methodology for UL channel generation for FDD systems agreed for FDD CSI enhancement in Rel-17 at RAN1#102-e.*

InterDigital

**Proposal 2: Study further the following aspects for UE-part 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**
* **Reporting contents/structure of UE-side monitoring metric and its associated feedback overhead**
* **NW-side monitoring with lower signaling overhead**

Vivo

1. **RAN1 concludes that both NW side monitoring and UE side monitoring should be supported for CSI compression.**
2. **RAN1 concludes that it is recommended to specify ground-truth CSI reporting via legacy codebook with potential configuration enhancement for NW side monitoring**
3. **RAN1 concludes that it is recommended to specify procedures for exchanging proxy model for approximating the quality of reconstructed model at UE side for UE-side monitoring.**

Apple

**Proposal 3: For CSI compression using two-sided model, for UE side performance, further study the NW implicitly transmit output CSI using precoded CSI-RS to the UE, and using hypothetical BLER as the performance metric.**

**Proposal 4: For CSI compression using two-sided model, for UE side performance, further study RLF/BFD like mechanism for UE initiated report.**

CATT

Proposal 14: In CSI compression using two-sided model use case, performance monitoring at NW-side can be prioritized and proxy model based performance monitoring is deprioritized.

China Telecom

***Observation 3: Whether to introduce proxy model needs to be justified.***

***Proposal 6: In CSI compression using two-sided model use case, deprioritize the study on UE-side monitoring based on the output-CSI transmitted from NW to UE.***

***Proposal 7: 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 reported by the UE.***

CMCC

***Proposal 9: 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.***

Fujistu

***Observation 22:***

* *For UE-side AI/ML model performance monitoring using a proxy model, the expectation of a simple structure and small size contradicts to the needs of a strong generalization capability for a proxy model to work well in various scenarios.*

***Observation 23:***

* *Using multiple proxy models for UE-side AI/ML model performance monitoring results in additional burden for model management, as well as potential additional overhead because of the assistance information required for choosing a right proxy model among multiple ones.*

***Observation 24:***

* *The proxy model used in the UE-side AI/ML model performance monitoring is also data-driven. And the performance of the proxy model should also be monitored regularly.*

***Observation 25:***

* *Only one SGCS is given by a proxy model, which may not be able to represent the performance of multiple CSI reconstruction models from multiple vendors.*

***Proposal 17:***

* *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 18:***

* *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 19:***

* *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 20:***

* *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 21:***

* *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 22:***

* *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.*

Xiaomi

***Proposal 3: It is necessary and feasible that performance monitoring by using intermediated KPIs or an existing CSI feedback scheme as the reference considering the following aspects:***

* ***Significant overhead reduction for quantization of the target CSI or output of CSI reconstruction via enhanced eType II codebook parameters***
* ***Affordable complexity for quantization of the target CSI or output of CSI reconstruction, which is similar to that of legacy eType II codebook.***
* ***Ensuring the robust of monitoring performance by using an existing CSI feedback scheme as the reference***

NEC

***Proposal 9: Support NW-side monitoring based on the target CSI with realistic channel estimation associated to the CSI report.***

***Proposal 10: For NW-side monitoring, the AI CSI and associated target CSI can be reported in the same reporting instance, or two separate reports.***

***Proposal 11: Support UE-side monitoring based on the output of the CSI reconstruction model.***

***Proposal 12: If the CSI reconstruction model at UE side is proven to be feasible, at least support UE-side monitoring*** ***based on the output of the CSI reconstruction model at UE side.***

Google

***Proposal 4: Do not support to use SGCS as the metric for ML performance monitoring.***

***Proposal 5: Support the hypothetical BLER as the metric for ML performance monitoring.***

***Proposal 6: Support the baseline for model performance monitoring based on the non-ML based CSI, i.e. the CSI based on existing codebook that the UE supports.***

* ***A model performance failure is identified if the hypothetical BLER measured based the ML based CSI and the CQI from the non-ML based CSI is above a threshold***
* ***ML based CSI compression should not mandate the UE to support eType2 codebook***

ZTE

***Proposal 16:*** *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 17:*** *In CSI compression using two-sided model use case, deprioritize the study on UE-side monitoring in Rel-19 study phase.*

Panasonic

**Observation 18: 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.**

ETRI

**Proposal 5: For AI/ML-based CSI compression using two-sided model, when the target CSI is Future slot(s), study method to align whether prediction and compression occur in separate steps or simultaneously between UE and NW, for inference and performance monitoring operations.**

Nokia

**Proposal 15: For NW-side performance monitoring in CSI compression using two-sided model use case, prioritize study of the framework in which the legacy non-AIML based CSI feedback mode is integrated, and fallback to the legacy non-AIML based CSI operation should be supported.**

Qualcomm Incorporated

***Proposal 5: Model monitoring solutions should not create a dependency between ML-based CSI feature and UE capability for eType-II or enhanced eType-II CSI feedback features. RAN1 should support a model monitoring solution that will work even for UEs without the capability to support enhanced eType-II based CSI feedback, or the capability to concurrently support eType-II like CSI feedback and ML-based CSI feedback.***

***Proposal 6: For model performance monitoring, RAN1 should support the UE-side monitoring method that directly outputs intermediate KPI at the UE side.***

***Proposal 7: For inter-vendor collaboration 3a/5a, conclude that the SGCS estimator can be developed by UE side with offline engineering.***

***Proposal 8: For inter-vendor collaboration 3b/5b, conclude that the SGCS estimator can be developed by NW and transferred from NW to UE. The unified signalling and framework used to transfer CSI generation model and CSI reconstruction model can be also used to transfer SGCS estimator.***

***Proposal 9: For UE side monitoring using SGCS estimator, study mechanisms with NW configuration and UE-initiated report of performance metric.***

Fraunhofer

***Proposal 2: For the two-sided model for the CSI feedback or CSI prediction use cases, the gNB monitors the performance of the AI model and detects possible faults based on the CSI report from the UE.***

## Discussion

|  |  |
| --- | --- |
| Companies | Views |
| Huawei | * Intermediate KPI or eventual KPI as starting point * Reporting mode: per sample or statistic reporting * Input / output distribution based method deprioritized * Proxy model based monitoring is level x * UE side monitoring w/ reconstructed CSI sent from NW * NW side monitoring w/ ground-truth reporting from UE |
| Intel | New report format for ground-truth CSI  SRS as a method for acquiring ground-truth for NW side monitoring |
| InterDigital | Study mechanisms of time-based or event-triggered reporting  Appropriate UE side monitoring metric and its reporting overhead  NW side monitoring with lower signaling overhead |
| vivo | Both NW side and UE side monitoring supported   * Ground-truth reporting w/ enhanced high-resolution codebook * Specify procedure for exchanging proxy model for approximating the quality of reconstructed CSI at UE side |
| Apple | UE side monitoring, NW sent reconstructed CSI using precoded CSI-RS, hypo-BLER as metric  RLF/BFD like mechanism for UE initiated report |
| CATT | NW side monitoring prioritized  Deprioritize proxy model at UE side |
| China Telecom | Deprioritize UE side monitoring based on output-CSI sent from NW.  Prioritize spec impact of ground-truth reporting for NW side monitoring |
| CMCC | NW side monitoring w/ ground-truth report  UE side monitoring based on reconstructed CSI from NW |
| Fujistu | * Proxy model is simple, so may have generalization issue. Multiple proxy model is complicated for model management. Proxy model needs monitoring as well. * Prioritize ground-truth reporting with high resolution codebook * Study options for monitoring NW side model using existing CSI feedback scheme * Study follow-up action, e.g., fallback to legacy codebook based CSI * Monitoring inactive model |
| Xiaomi | Intermediate KPI using existing CSI feedback scheme as reference considering overhead and complexity |
| NEC | NW side monitoring using ground-truth reporting, in one report with nominal inference, or separate reports  UE side monitoring based reconstructed CSI (if decoder is available) |
| Google | * Do not use SGCS, use hypo BLER instead * Monitoring based on existing codebook, but not UE not mandate to support eT2 * Model failure = hypo BLER measured using ML CSI and CQI from non-ML CSI is above a threshold |
| ZTE | NW side monitoring w/ ground-truth reporting  Deprioritize UE side monitoring |
| Panasonic | NW side monitoring w/ ground-truth reporting  UE side monitoring w/ UE side reconstruction model |
| Nokia | Prioritize monitoring integrating non-ML CSI and fallback mode to the non-ML CSI |
| QC | No dependency between monitoring and eT2  Support proxy based UE side monitoring (SGCS estimator) – can be developed by UE side offline engineering or NW side with parameter / model transfer  Study NW configuration of monitoring or UE initiated monitoring report |
| Fraunhofer | gNB monitoring and detect possible faults based on CSI report from UE |

A good number of companies supporting NW side monitoring w/ ground-truth reporting (11)

* High-resolution codebook-based CSI feedback
* SRS

Concern raised: signaling overhead, latency, dependency w/ UE capability of supporting eT2 or its enhanced version.

Also, a good number of companies interested in UE side monitoring (8)

* Reconstructed CSI indication from NW to UE
* UE side running inference using CSI reconstruction model
* Proxy model for SGCS estimation
* Hypo BLER measured by precoded CSI-RS
* Hypo BLER measured by codebook based CSI (w/o dependency on eT2)

For these UE-side monitoring approaches, reconstructed CSI indication from NW to UE incurs large DL overhead, UE side running CSI reconstruction model may incur additional cost of timeline and CSI processing requirements.

For proxy model used for SGCS estimation, there are concerns on additional effort of model management and generalization. One company (Qualcomm) provides evaluation on monitoring accuracy and generalization aspects. Hypothetical BLER methods are proposed by Google and Apple, more study is needed for monitoring accuracy.

From FL perspective, NW side monitoring, UE side monitoring with reconstructed CSI indication from NW, and UE side running reconstruction model or proxy model have been studied in Rel-18, it would be helpful for progress if we can conclude something. For other options of UE side monitoring, they are good candidates that address signaling overhead, and UE capability / complexity issue, thus it is beneficial to give more attention for further study.

Besides, reporting mechanism discussion is also provided by companies, possible options include UE initiated CSI or configured / time-based; also, some companies mentions how existing non-ML CSF integrated and considered as fallback. These aspects can be further discussed.

### Monitoring via ground truth reporting

Proposal 41a:

**Further study following monitoring options in Rel-19, considering how to reduce signaling overhead and latency, and how to relax UE capability on reporting target CSI using eT2-like high-resolution codebook, while maintaining a good monitoring performance**

* **NW side monitoring with target CSI reported from UE side.**
* **UE side monitoring with output of CSI reconstruction model indicated from NW side.**

|  |  |
| --- | --- |
| *Support / Can accept* | Lenovo, Futurewei |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Proxy model is widely studied in Rel-18.  **Further study following monitoring options in Rel-19, considering how to reduce signaling overhead and latency, and how to relax UE capability on reporting target CSI using eT2-like high-resolution codebook, while maintaining a good monitoring performance**   * **NW side monitoring with target CSI reported from UE side.** * **UE side monitoring with output of CSI reconstruction model or Proxy model indicated from NW side.** |
| Huawei, HiSilicon | Support using eT2-like high-resolution codebook for monitoring. But there is one issue:  Referring the LS reply (R1-2310681) from RAN1 to RAN2, the latency requirement of monitoring is near real time, which is tens of ms to seconds. Thus, the signalling interaction for monitoring should be between gNB and UE, rather than to a UE side OTT server where the latency is not guaranteed.      Therefore, “UE side” is changed to “UE”.  In addition, “output of CSI reconstruction model” is replaced with “reconstructed target CSI” to be aligned with previous agreement.  **Further study following monitoring options in Rel-19, considering how to reduce signaling overhead and latency, and how to relax UE capability on reporting target CSI using eT2-like high-resolution codebook, while maintaining a good monitoring performance**   * **NW side monitoring with target CSI reported from UE ~~side~~.** * **UE ~~side~~ monitoring with reconstructed target CSI ~~output of CSI reconstruction model~~ indicated from NW side.** |
| Panasonic | We support vivo’s update. |
| Fujitsu | We are generally fine with the original version from FL. |
| LG | Fine with further study |
| CATT | We agree with Huawei’s proposal. We are not in favour of the proxy model study. Given current workload and limited time, further enlarging the study scope for proxy model id not preferred. |

### Monitoring via running CSI reconstruction model

Proposal 42a:

**Further study UE side monitoring with running inference using CSI reconstruction model, considering UE side complexity, e.g., timeline, CSI processing criteria (CPU occupation, active resource/port occupation)**

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Lenovo | We suggest to combine it with the next proposal  If the preference is to keep proposal 42a and 43a seperare we suggest to add the following to this proposal.  **Further study UE side monitoring with running inference using the actual or proxy model for the CSI reconstruction model, considering UE side complexity, e.g., timeline, CSI processing criteria (CPU occupation, active resource/port occupation)** |
| vivo | Fine with the version from Lenovo. |
| Huawei, HiSilicon | Disagree. Introducing proxy model will enlarge the NW side model management burden; in addition, how to monitor a proxy model is not clear. |
| Panasonic | We support Lenovo’s update. |
| ZTE | Agree with Huawei’s comments. |
| Fujitsu | Don’t support.  Basically agree with Huawei. The reliability of UE-side monitoring using proxy model is questionable. The performance of proxy model should be monitored and is not clear. This also increases the UE side burden. |
| LG | Fine with the proposal |
| CATT | Agree with Huawei, how to monitor a proxy model is not clear. |

### Monitoring via running CSI reconstruction model

Proposal 43a:

**Further study following UE side monitoring methods, considering UE side complexity, monitoring performance and reporting overhead, including evaluation of monitoring performance and generalization ability (if relevant).**

* **Proxy model for SGCS estimation**
* **Hypo BLER measured by precoded CSI-RS**
* **Hypo BLER measured by codebook based CSI (w/o dependency on eT2)**

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Lenovo | Add the following bullet   * **Using the actual or proxy model of the CSI reconstruction model**   If the preference is to keep proposal 42a and 43a seperare we suggest to add the following to this proposal as well.  **considering UE side complexity, e.g., timeline, CSI processing criteria (CPU occupation, active resource/port occupation)** |
| Huawei, HiSilicon | The monitoring performance and generalization ability for UE side proxy model has been evaluated in R18, and observations also capture relevant results. Can I clarify the purpose of this proposal: is it encouraging companies to continue the evaluation and collect more results? |
| ZTE | There is no sufficient evidence on the feasibility of all the cases. It can be deferred to discuss later. |
| Fujitsu | Don’t support.  This could be further discussed if using proxy model have been agreed. |

### 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.

Huawei

***Proposal 15: 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 16: The down selection of the model pairing options can be discussed in Rel-19 after other aspects are clearer, e.g., model identification, training collaboration types.***

***Proposal 17: 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 18: 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.***

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.***

Apple

**Proposal 1: For time-frequency-spatial domain CSI compression, the following potential specification impact are proposed:**

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

**Proposal 2: For time-frequency-spatial domain CSI compression, flexible CSI report configuration to support different cases should be studied.**

CATT

Proposal 16: In CSI compression using two-sided model use case, legacy CSI reporting principles is reused as much as possible.

Proposal 17: 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 18: 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.**

China Telecom

***Proposal 5: For CSI compression sub use case, the pairing information should be included in the process of functionality/model identification.***

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 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 #4: Consider the method on the rank adaptation based on the validity check of layer(s) for a given RI.**

**Proposal #5: 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.**

Fujistu

***Proposal 11:***

* *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 12:***

* *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 13:***

* *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 14:***

* *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 15:***

* *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 16:***

* *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.*

Xiaomi

***Proposal 4: 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 5: 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 6: 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 7: 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.***

NEC

***Proposal 2: For Case 2, CSI buffer reset should be supported to address UCI loss and rank adaption. And the definition, determination or indication of the reset value need to be further studied.***

***Proposal 3: For Case 2, further study*** ***effective availability of past CSI information over time.***

***Proposal 13: If the*** ***CSI reconstruction part at UE side is proven to be feasible, at least support Option 2a for CQI determination in CSI report. If not, support Option 1a/1b.***

***Proposal 14: 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.***

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 15:*** *For CQI determination, at least prioritize the specification impact discussions on Option 1a, Option 1b.*

Panasonic

**Observation 6: For complexity comparison to study the scalability of rank > 1 solutions, the total complexity with multiple models should be taken into account.**

**Observation 7: 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.**

ETRI

**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.**

Nokia

Proposal 12: RAN1 to focus on the evaluation of Options 1a and 2a-1 for CQI calculation, also considering proposals for Options 1b and 2a-2.

Proposal 13: RAN1 to study the feedback of CQI for different rank hypotheses.

**Proposal 14: RAN1 to study the specification effect of layer common, layer specific, rank common, and rank specific architectures to determine how specifications affect which architectures are supported.**

Qualcomm Incorporated

***Proposal 10: 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.***

***Proposal 11: 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***

***Proposal 12: Proposal: 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.***

* 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.

***Proposal 13: 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-5: Consider the possibility of using model-ID based standardized model for the same architecture with different configurations.**

**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-7: Consider using model ID based identification for ensuring proper training between UE sided model and NW sided model**

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

## Discussion

|  |  |
| --- | --- |
| Companies | Views |
| Huawei | Standardization of VQ/SQ configuration and dictionary  Downselection model paring options after model identification / training collaboration types are clear  CQI option 1  CSI priority rule  CPU  UCI design |
| TCL | AI/ML specific CSI-RS config, fields in CSI report, dedicated report quantity  CSI Priority rule  How to describe UE capabilities of running inference |
| Apple | For TSF, study DCI rest memory of UCI and UCI reTx  For TSF, flexible CSI config to support various use cases |
| CATT | For CQI, consider sub-options of option 1 |
| China Telecom | Paring information included in functionality/model identification |
| 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 |
| Fujistu | Model paring: study UE-initiated and NW-initiated approaches  Paring information could be model ID (global)  Downeselect rank/layer specific/common model  Deprioritize CQI option2 |
| Xiaomi | CSI priority rule  CSI omission and UCI part 2 partition  UCI packing of CSI in multiple predicted slots |
| NEC | CSI buffer rest to address UCI loss and rank adaptation  Support CQI option 2a, 1a/1b  Down-select paring information from encoder ID, model ID, pair ID |
| Google | CSI priority rule  CPU occupancy  Hybrid AI-CSF and non-AI CSF (depending on RI) |
| ZTE | Support CQI option 1a/1b |
| ETRI | Study method to mitigate inconsistency of past CSI |
| Nokia | CQI option 1a, 1b, 2a-1, 2a-2  CQI feedback for different rank hypo  Study specification effort of layer/rank-common/specific model |
| QC | Conclude CQI option 2a can be used for higher cost of time and complexity  Further study CQI option 1b considering intermediate KPI  Study layer-common model structure w/ potentially layer-specific or common parameters  Study the quantization alignment levels |

### Handling of misalignment of historical CSI

From the summary table, we can see that some companies propose to study how to address misalignment of past CSI for TSF case. In legacy, all CSI feedback schemes are self-contained, the dependency of previous CSI report at UE side and gNB side is an important aspect to address for TSF case. Thus, it may be helpful to trigger some study for this aspect.

Proposal 51a:

For temporal domain aspects Case 2 / 4, study mechanisms that manage historical CSI at two sides and solutions that address misalignment of historical CSI used at UE side and NW side due to UCI loss or dynamic rank reporting. Following can be considered as examples

* UCI memory / buffer reset
* (Re)transmission of historical CSI
* Other options are not precluded

|  |  |
| --- | --- |
| *Support / Can accept* | Huawei, HiSilicon, LG |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | May not hurry to study these aspects before conclusion of TSF cases. |
| Panasonic | We are OK to study these mechanisms/solutions, but we tend to agree to vivo that we need conclusions of TSF cases before studying these mechanisms/solutions. |
| ETRI | We support on the direction of the proposal. One question for clarification on the first buller is that UCI memory/buffer reset on UE? If then, in our view, UE can also generate appropriate UCI based on the awareness of UCI loss at UE, without mandatory UCI memory resetting on UE. |
| Fujitsu | We are ok to discuss this issue and suggest the following update to include one more option.  *For temporal domain aspects Case 2 / 4, study mechanisms that manage historical CSI at two sides and solutions that address misalignment of historical CSI used at UE side and NW side due to UCI loss or dynamic rank reporting. Following can be considered as examples*   * *UCI memory / buffer reset* * *(Re)transmission of historical CSI* * *Switch between Case 2/4 and Case 0 or non-AI/ML method* * *Other options are not precluded* |

### CQI determination

Several options for CQI determination were listed in TR 38.843:

|  |
| --- |
| For CQI determination in CSI report, if CQI in CSI report is configured.  - 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  - 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 reconstruction model inference with potential adjustment  - Note: CSI reconstruction part at the UE can be different comparing to the actual CSI reconstruction part used at the NW.  - Option 2b: CQI is calculated using two stage approach, UE derive CQI using precoded CSI-RS transmitted with a reconstructed precoder.  - Notes: feasibility of different options should be evaluated. Gap analyses between the UE side CQI calculation results and the NW side results, as well as the impact on the scheduling performance should be evaluated. Complexity of CQI calculation needs to be evaluated, including the computing complexity and potential RS/signalling overhead. |

Some companies propose their preference on option 2a and 1a/1b. Thus, it would be good to draw some conclusion in this release (if possible) and encourage companies to provide more detailed study for option 1a/1b, e.g., the detailed adjustment examples and its accuracy.

Conclude that option 2a can be considered if decoder available, but with additional requirements of timeline, CPU, active port/resource occupation.

Further study details, feasibility, performance of CQI option 1a/1b

Proposal 52a:

* Conclude that option 2a can be considered if CSI reconstruction model or its reference model is available at UE, but with additional requirements of timeline, CPU, active port/resource occupation.
* Further study details, feasibility, performance of CQI option 1a/1b.

|  |  |
| --- | --- |
| *Support / Can accept* | Futurewei, ETRI |
| *Object / Have a concern* | Huawei, HiSilicon |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| vivo | Do not support.  We believe Option1b might be a better solution. |
| Huawei, HiSilicon | We think the most simple and feasible solution is Option 1a/1b. The NW side can adopt OLLA to adjust the MCS, which is broadly used in legacy. |
| Fujitsu | Share similar view as vivo and Huawei. |
| LG | Generally fine with the proposal. |

### Model pairing & quantization alignment

Some companies propose to down-select among model paring information options, FL thinks it would be good to combine with inter-vendor collaboration options. Also, for quantization alignment and rank/layer aspects of model design / structure, FL would encourage companies to propose detailed and concrete proposals so as to help draw conclusion and identify necessary spec impacts.

Study and down-select model paring information and model identification procedure considering inter-vendor collaboration options

Study spec effort of model scalability of layers and rank, downselect from layer/rank-common/specific model structure

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

### Others

Please provide any other comments for this section.

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

# 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.

Samsung

**Proposal#9: Study the impact of input pre-processing (dimensionality reduction) on performance and model complexity.**

**Proposal#10: In Angle-delay (W2)-domain CSI compression, study the impact of the number of SD/FD basis vectors for performance-complexity tradeoff.**

## Discussion

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Intel | In the TR from Rel-18 study on AI/ML for PHY (TR 38.843, Section 6.2.2.8) it was mentioned that the complexity between AI/ML and non-AI/ML benchmark is not compared. We think it is important to have some (at least approximate, in some range) understanding of PMI search and PMI reconstruction complexity for non-AI/ML codebook-based CSI to have a meaningful RAN1 conclusion in this release. |
|  |  |

# 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

To be provided at the end of the meeting

# 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

# 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 #117, Apr. 2024.
5. R1-2402026 Discussion on AI/ML for CSI compression Huawei, HiSilicon
6. R1-2402053 Discussion on improving trade-off between performance and complexity/overhead for AI/ML-based temporal-domain CSI feedback compression. FUTUREWEI
7. R1-2402096 Discussion on AIML for CSI compression Spreadtrum Communications
8. R1-2402147 AI/ML for CSI compression Intel Corporation
9. R1-2402233 Discussion on CSI compression vivo
10. R1-2402266 Discussion on study for AI/ML CSI compression ZTE
11. R1-2402279 AI/ML based CSI Compression Google
12. R1-2402319 Additional study on AI/ML-based CSI compression OPPO
13. R1-2402369 Additional study on AI/ML-based CSI compression CATT
14. R1-2402455 Discussion for further study on AI/ML-based CSI compression Samsung
15. R1-2402495 AI/ML for CSI compression Ericsson
16. R1-2402506 Discussion on AI/ML-based CSI compression China Telecom
17. R1-2402526 Discussion on CSI compression for AI/ML BJTU
18. R1-2402556 Discussion on AI/ML for CSI compression CMCC
19. R1-2402630 Study on CSI compression LG Electronics
20. R1-2402652 Discussion on two-sided AI/ML model based CSI compression Xiaomi
21. R1-2402750 Discussion on AI/ML for CSI compression Panasonic
22. R1-2402766 Discussion on CSI compression NEC
23. R1-2402789 Discussion on CSI compression with AI/ML Fujitsu
24. R1-2402843 Discussion on AI/ML-based CSI compression InterDigital, Inc.
25. R1-2402849 Addtional study on AI-enabled CSI compression NVIDIA
26. R1-2402872 Discussion on AI based CSI compression Apple
27. R1-2402921 On AI/ML for CSI compression Lenovo
28. R1-2402960 Discussion on CSI compression Sony
29. R1-2402999 AI/ML for CSI Compression Nokia
30. R1-2403013 Discussion on AI/ML for CSI compression ETRI
31. R1-2403054 Discussion on AI/ML for CSI Compression CEWiT
32. R1-2403076 Additional Study on AI/ML for CSI Compression MediaTek
33. R1-2403100 Discussion on AI/ML for CSI compression SK Telecom
34. R1-2403147 Discussion on AI/ML for CSI compression AT&T
35. R1-2403158 Discussions on AI/ML for CSI feedback CAICT
36. R1-2403185 Additional study on CSI compression Qualcomm Incorporated
37. R1-2403235 Discussion on AI/ML for CSI compression NTT DOCOMO, INC.
38. R1-2403279 AI/ML based CSI compression ITL
39. R1-2403336 Discussion on the AI/ML for CSI Compression Fraunhofer IIS, Fraunhofer HHI
40. R1-2403380 Discussion on study of AI/ML for CSI compression IIT Kanpur, Indian Institute of Tech (M)
41. R1-2403381 Discussion on Additional Study of AI/ML for CSI Compression Indian Institute of Tech (M), IIT Kanpur