**3****GPP TSG RAN WG1 #116-bis R1-240xxxx**

**Changsha, China, April 15th – April 19th, 2024**

**Agenda item:** 9.1.3.2

**Source:** Moderator (Qualcomm)

**Title:** FL summary for 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.comwangjf20@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 WangHaruhi Echigo | wangx@docomolabs-beijing.com.cnharuhi.echigo.fw@nttdocomo.com |
| ETRI | Anseok Lee | alee@etri.re.kr |
| Ericsson | Jingya LiSiva MuruganathanJianwei Zhang | Jingya.li@ericsson.comSiva.muruganathan@ericsson.comJianwei.zhang@ericsson.com |
| ZTE | Lun LiXingguang Wei | li.lun1@zte.com.cnwei.xingguang@zte.com.cn |
| IIT Kanpur | Shyam Vijay GadhaiAbhishek Kumar Singh | svgadhai@iitk.ac.inabhishekks@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 | LiuzhengxuanLiumin | liuzhengxuan@xiaomi.comliumin10@xiaomi.com |
| AT&T | Isfar TariqSalam Akoum | Isfar.tariq@att.comSalam.akoum@att.com |
| CMCC | Yuhua CaoYi ZhengDan Song | caoyuhua@chinamobile.comzhengyi@chinamobile.comsongdan@chinamobile.com |
| NVIDIA | Xingqin Lin | xingqinl@nvidia.com  |
| MediaTek | Pedram Kheirkhah Sangdeh | Pedram.kheirkhah@mediatek.com |
| Futurewei | Baoling S Sheen | bsheen@futurewei.com  |
| TCL | Yunsheng Kuang | yunsheng.kuang@tcl.com |
| Huawei, HiSilicon | Yuan Li | liyuan3@huawei.com |
| CEWiT(TSDSI) | Shiv ShankarDhivagar Baskaran | shivshankar@cewit.org.indhivagar.b@cewit.org.in |

# 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

From the submitted contributions, proposals related to temporal domain aspects of CSI compression are summarized below.

Agreement

***For the evaluation of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Release 19, adopt the following categorization for study:***

|  |  |  |  |
| --- | --- | --- | --- |
| **Case** | **Target CSI slot(s)** | **Whether the UE uses past CSI information** | **Whether the network uses past CSI information** |
| 0 | Present slot | No | No |
| **1** | Present slot | Yes | No |
| **2** | Present slot | Yes | Yes |
| **3** | Future slot(s) | Yes | No |
| **4** | Future slot(s) | Yes | Yes |
| **5** | Present slot | No | Yes |

***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.***

Huawei

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 2: 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 (taking time instance t=2 for example):
	+ Model input: original CSI of the current slot ($V\_{2}$) and accumulated CSI information from the last time instance ($W\_{1}$).
	+ Model output: CSI feedback of the current slot ($V\_{Q2}$) and accumulated CSI information for the next time instance ($W\_{2}$).
* CSI reconstruction part (taking time instance t=2 for example):
	+ Model input: CSI feedback of the current slot ($V\_{Q2}$) and accumulated CSI information from the last time instance ($W\_{1}^{'}$).
	+ Model output: recovery CSI of the current slot ($\tilde{V}\_{2}$) and accumulated CSI information for the next time instance ($W\_{2}^{'}$).
* Note: after inference, the accumulated CSI information at the CSI generation part and CSI reconstruction part are updated from $W\_{1}$ to $W\_{2}$ and from $W\_{1}^{'}$ to $W\_{2}^{'}$, respectively. The update of the accumulated CSI information does not impact the weights of the models.

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.

Samsung

***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***

* Option1: Past CSI information generated by the UE-part and/or network-part of two-sided model
* Option 2: Information on SD/FD basis vectors as past CSI information with angle-delay (W2) 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 value 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*** $N\_{4}\geq 1$ ***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: Among the identified six categories for AI/ML-based CSI compression using two-sided model, for Case 5, consider***

* ***Option1: eigenvectors-based target CSI (right eigenvectors and/or left-eignvectors of full channel matrices)***
* ***Option 2: raw channel matrices-based target CSI***

ZTE

***Proposal 1:*** *For temporal domain CSI compression, further study and evaluate the performance on Case 2 and Case 3 as a starting point.*

***Proposal 2:*** *For CSI compression plus prediction sub-use case in Case 3, legacy CSI prediction plus AI CSI compression should be prioritized to study and evaluate the performance.*

***Proposal 3:*** *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 4: For temporal domain CSI compression Case 2, further study the performance impact resulting from the aperiodic CSI feedback.***

OPPO:

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

* ***Study Case 2 and Case 5 without CSI prediction in high priority***
* ***Study Case 3 and Case 4 with CSI prediction in low priority***
* ***Use Case 0 as additional benchmark for performance comparison***
* ***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: suggest to evaluate and discuss non-ideal UCI feedback after the performance gain of SFT-domain CSI compression is verified.***

***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***

CATT

***Proposal 1: For the evaluation of 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 the evaluation of 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 the evaluation of 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 the evaluation of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Rel-19, considering the following sub-cases for Case 4:***

* ***Case 4-1: target CSI is the CSI for a future instant***
	+ ***Case 4-1a: CSI generation part without past CSI information (CSI prediction+ Case 5)***
	+ ***Case 4-1b: CSI generation part with past CSI information (CSI prediction+ Case 2)***
* ***Case 4-2: target CSI is the CSI for multiple future instants***
	+ ***Case 4-2a: CSI generation part without past CSI information (CSI prediction+ Case 5 multiple slots extension)***
	+ ***Case 4-2b: CSI generation part with past CSI information (CSI prediction+ Case 2 multiple slots extension)***

***Proposal 5: For the evaluation of 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.***

***Proposal 6: For the evaluation result template to capture the results of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Rel-19, similar table as table 1 in Rel-18 CSI compression can be considered,****,*

* ***The focus should be on validating performance enhancement for temporal domain aspects of AI/ML-based CSI compression and associating temporal information configuration regime that facilitates such performance enhancement over Rel-18 spatial temporal domain CSI compression.***
* ***Companies should report UE speed, UE distribution, realistic channel estimation error modeling, and observation/prediction window configuration, if applicable.***
* ***Evaluation result template to capture model scalability, multi-vendor joint training or separate training should be deprioritized.***

Futurewei

***Proposal 2: For temporal-domain CSI compression using two-sided model sub use case, consider incorporating the following attributes into the performance evaluation template(s) adopted in Rel-18:***

* ***General assumptions***
	+ ***CSI-RS configuration adopted: periodic or aperiodic, and corresponding periodicity where applicable.***
	+ ***CSI report periodicity***
* ***Under the assumptions related to modelling:***
	+ ***Temporal-domain CSI compression case***
	+ ***Observation window (number/distance)***
	+ ***Prediction window(number/distance)***
* ***Under the benchmark or baseline***
	+ ***UE distribution***
* ***Under the evaluation section***
	+ ***CSI feedback overhead rate***

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

China Telecom

***Proposal 1: To further discuss the potential alternatives for the legacy CSI prediction plus AI CSI compression and the AI CSI prediction plus AI CSI compression.***

CMCC

***Proposal 1: At least one case of “present slot” (e.g. case 2), and one case of “future slot” (e.g. case 3 or 4) can be focused during future meetings.***

NEC

***Proposal 1: Study to use model compression to reduce the complexity of the AI/ML model(s) at least for Case 1~4.***

Spreadtrum

***Proposal 1: Prioritize evaluating case 1, case 2 and case 5 for SFT-domain CSI compression.***

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 study the performance impact due to unavailable past CSI information for some layers resulting from rank adaption.

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.

NVIDIA

Apple

Proposal 3: 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.

Lenovo

Proposal 8: Prioritize Case 0, Case 3 and Case 5 for temporal domain aspects of AI/ML-based CSI compression using two-sided model, whereas Case 1, Case 2 and Case 4 are deprioritized.

Proposal 9: Strive to unify the CSI framework across the two agendas for AI/ML study of CSI feedback compression and CSI prediction enhancements.

Sony

Nokia

Proposal 1: Regarding SFT-based CSI compression, prioritize on 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.

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.

CEWiT

Proposal-1: Companies to report the number of instances which can be batch input for temporal compression of CSI.

Proposal-2: Consider the effect of CSI periodicity for a given UE speed while considering the batches of CSI input to the model for AI/ML based SFT-CSI compression.

Proposal-3: 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.

MediaTek

SK Telecom

Proposal 1 For the evaluation of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Release 19, Case 3 and 4 (i.e., Target CSI slot(s) = Future slot (s)) are prioritized than others.

AT&T

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

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

CAICT

Qualcomm

Proposal 3: For capturing the evaluations results on temporal domain compression case 2, Rel-18 results Table-1 can be used as a starting point with the following modifications:

* Add the time domain assumptions (e.g., UE speed, CSI-RS periodicity, CSI feedback periodicity, …, etc.).
* Add the descriptions of how historical CSI measurement and CSI report are used at UE side and/or gNB side (e.g., number/instance of historical measurement and CSI reports).

Proposal 4: For capturing the evaluations results on joint prediction and compression (case 3), Rel-18 results Table-1 can be used as a starting point with the following modifications:

* Add the time domain assumptions (e.g., UE speed, CSI-RS periodicity, CSI feedback periodicity, observation window, prediction window, …, etc.).
* Add the descriptions of the prediction algorithms for the baseline and the AI/ML algorithms (e.g., input type, output type, and prediction method.).

Proposal 5: Study techniques and potential specification impact to enable the use of localized models to achieve the associated improvement in the performance-complexity tradeoff.

Proposal 6: Capture the following aspects in results table for localized models:

* Description of the localized region that is used to train / develop the model
* Benchmark scheme contains the model trained under global dataset

NTT DOCOMO

Proposal 1

* Prioritize Case 3 and Case 4 for Rel. 19 study of CSI compression with temporal domain aspects, as Case 3 and Case 4 can increase the scheduling flexibility by reducing the number of CSI reporting occasions.

Proposal 3

* For the CSI compression with temporal domain aspects, further study the generalization over the following aspects,
* UE speed,
* CSI-RS periodicity or different CSI-RS burst configurations.
* Study the generalization over multiple aspects as the following combinations,
* Deployment scenarios (indoor/outdoor scenarios) + carrier frequency.
* Deployment scenarios + frequency granularity.
* Deployment scenarios + antenna port number.
* Deployment scenarios + carrier frequency + frequency granularity.
* Deployment scenarios + carrier frequency + antenna port number.

ITL

Proposal 1: It is proposed to consider prediction-based methods (Case 3 and 4) only when prediction accuracy can be ensured

Fraunhofer IIS, Fraunhofer HHI

IIT KANPUR, INDIAN INSTITUTE OF TECHNOLOGY MADRAS (IITM)

Indian Institute of Technology Madras (IITM), IIT Kanpur

## Discussion

Below is a quick summary of companies’ views and discussion topics.

|  |  |
| --- | --- |
| Company | Views |
| Huawei | * Prioritize case 2 / 3
* Capture input/output in EVM
* Modeling of UCI missing, how to handle it
* LCM of 1 or 2 models for joint prediction and compression
 |
| Samsung | * List input/output options for case 2, SD/FD basis can be one option of past CSI
* List options for case 3 / 4 (SFT, angle-delay-time, angle-delay-doppler)
* List options for case 5 (eigen-vectors or raw channel)
 |
| ZTE | * Prioritize case 2 / 3
* Capture input/output in EVM for case 3
* Impact of A-CSI-RS for case 2
 |
| OPPO | * Prioritize case 2 / 5, deprioritize prediction case 3 / 4, no further case 1
* Non-ideal UCI
 |
| CATT | * Prioritize case 2, separate prediction compression for case 3 / 4
* Case 3 / 4 is based on separate prediction and compression. Expand case 3 / 4 depending on whether past CSI is used in compression
* Results table: capture time-domain behavior, model scalability, inter-vendor aspects
 |
| Futurewei | * Results table: time domain behavior, UE distribution, CSI feedback rate
 |
| LGE | * Sync of past CSI at the two sides, how to handle non-ideal UCI
* How to report historical CSI for monitoring
 |
| China Telecom | * Legacy CSI prediction + AI compression or AI prediction + AI compression
 |
| CMCC | * Prioritize case 2 + case 3 or 4
 |
| Spreadtrum | * Prioritize case 1 / 2 / 5 (no prediction)
 |
| Fujistu | * Handling UCI loss
* Unavailable historical CSI due to rank adaptation
 |
| InterDigital | * Increased complexity of TSF over SF
 |
| Apple | * Non ideal UCI handling, e.g., DCI reset, UCI reTx
 |
| Lenovo | * Prioritize case 0 / 3 / 5
 |
| Nokia | * Prioritize case 2 / 4
* Study time coherence aspects, e.g., various feedback periods, speeds, scenario, length of time sequence
* Non-ideal UCI
 |
| ETRI | * Prioritize case 2 / 4
* Study inconsistency of the availability of past CSI information
 |
| CEWiT | * For case 3 / 4, study joint and separate methods for compression and prediction
 |
| SK telecom | * Prioritize case 3 / 4
 |
| ATT | * Deprioritize joint prediction and compression, only consider UE performing prediction
 |
| Qualcomm | * Results table: time domain behaviors, description of historical CSI usage, prediction algorithm
 |
| DCM | * Prioritize case 3 / 4
* Generalization scenarios of speed, periodicity
* Generalization scenarios + frequency granularity + antenna port number
 |

### Prioritization of Cases

The table below is a summary of companies’ views regarding prioritization.

|  |  |  |
| --- | --- | --- |
|  | Prioritize | deprioritize |
| Case 0 | Lenovo |  |
| Case 1 | Spreadtrum | Lenovo |
| Case 2 (8) | Huawei, ZTE, OPPO, CATT, CMCC, spreadtrum, Nokia, ETRI | Lenovo |
| Case 3 (7) | Huawei, ZTE, CATT, CMCC, Lenovo, SK telecom, DCM | OPPO, ATT |
| Case 4 (6) | CATT, CMCC, Nokia, ETRI, SK telecom, DCM | OPPO, ATT, Lenovo |
| Case 5 | OPPO, spreadtrum, lenovo |  |

In general, historical CSI provides time correlation information, such information can be used to benefit compression efficiency and/or predication accuracy.

For compression cases (case 1 / 2 / 5), since two-sided model is used, it is expected that the time correlation information can be leveraged at both sides to maximize the performance. Thus, case 2 is more favoured than case 1 and 5. This is also mentioned by many companies including Huawei, ZTE, vivo, CATT, Nokia. Performance gain has been shown by Huawei, ZTE, vivo, CATT, Nokia, Xiaomi, Spreadtrum, Nokia, ETRI, Qualcomm.

For prediction + compression use cases (case 3 / 4), one can use single model to perform these two functions jointly or use separate model to perform these two functions in sequential manner. Hence, it is natural to only use historical CSI at UE side to assist CSI prediction, or use it at both sides to enhance both prediction accuracy and compression efficiency. Some evaluation results have been captured in the contributions of Huawei, Ericsson, vivo, ZTE, Qualcomm, etc. There are also voices arguing that mixing prediction and compression in case 3 / 4 complicates the study, but given the performance gain already shown by the aforementioned companies, it is fair to proceed with these cases.

Proposal 1a:

For the evaluation of temporal domain aspects of AI/ML-based CSI compression using two-sided model in Release 19, prioritize Cases 2 / 3 / 4 for further study.

Note: Companies can still provide evaluation results and analysis for Cases 1 and 5.

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Futurewei | We don’t think we need to prioritize the 5 cases till we have some observations from results submitted by companies. We also think clarification for Case 1 is needed as some companies think it is not clear the difference between Case 0 and Case 1. Based on our understanding, Case 1 may include a flavor/variation that UE will compress and send together current CSI and historical CSIs in the CSI report. In this variation, gNB implicitly has past CSI information when performing the reconstruction even if it doesn’t explicitly use past CSI information. |
| Huawei, HiSilicon | OK in general. One clarification question to Case 4: as the target CSI slot already provides the future CSI information, why does NW still need to use the historic/past CSI information? Does it mean NW also performs CSI prediction? |

### Non-ideal UCI feedback

Several companies discuss non-ideal UCI feedback, proposed solutions, and brought some evaluation results for temporal domain CSI compression Cases 2 and 4. Some examples of proposed solutions are:

* Two-step monitoring mechanisms to identify whether the issue lies in UCI loss or deployment scenario
* Past CSI reporting via NW-triggered signalling
* UCI retransmission
* DCI based reset memory

Also, it may be good to have some suggestions how to evaluate non-ideal UCI feedback. This is discussed in the “Results template” proposal.

Proposal 2a:

Proposal:

For temporal domain CSI compression Cases 2 and 4, study mechanisms to detect and handle the misalignment of past CSI information used at UE side and NW side due to UCI loss or rank adaptation.

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| New H3C  | OK in general |
| Huawei, HiSilicon | One question to “rank adaptation”: as the past CSI information is accumulated for per layer basis, changing the rank value seems no impact to the synchronization of the past CSI at both sides? |
|  |  |
|  |  |

### Further details for each Case

There are some proposals to expand and capture detailed options of each case (e.g., input/output or their types). However, FL thinks that it may be hard to align them and may bring little benefit to progress. Thus, they can be left to companies’ choices in their evaluations.

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

### Results template for Cases 1/2/5

Proposal 3a:

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
	+ CSI scheduling delay
* 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)
* 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)
* Modelling of spatial consistency
* Modelling of UCI loss (if applicable)
	+ E.g., 10% UCI loss, etc
	+ E.g., 1st UCI is received without error, 10% loss for subsequent UCIs, etc
* Modelling of realistic channel estimation
* Modelling of phase discontinuity

|  |  |
| --- | --- |
| *Support / Can accept* | Futurewei (with comments) |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Futurewei | To understand the gains over the BL, i.e., spatial-frequency CSI compression, we suggest including model complexity for both the temporal-domain CSI compression model and compared BL model.  |
| Huawei, HiSilicon | 1) “CSI scheduling delay” seems not needed – it does not 2) For “Methods to handle UCI loss”, it is better to provide some guidance for simulation cases, e.g., CSI buffer reset, CSI retransmission, etc.3) “spatial consistency” is not mandatory for Case 1/2/5. Changed to “whether/how spatial consistency is modelled”4) For “Modelling of UCI loss”, it is better to align the modelling in a separate proposal. BTW, can FL clarify the meaning of “1st UCI is received without error, 10% loss for subsequent UCIs”?5) Case 1/2/5 does not involve CSI prediction, so phase continuity and realistic channel modelling does not impact the performance over benchmark.* Temporal domain CSI setting
	+ CSI feedback periodicity
	+ CSI-RS periodicity
	+ ~~CSI scheduling delay~~
* 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)
* 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 ~~Modelling of~~ spatial consistency is modeled
* Modelling of UCI loss (if applicable)
	+ E.g., 10% UCI loss, etc
	+ E.g., 1st UCI is received without error, 10% loss for subsequent UCIs, etc
* ~~Modelling of realistic channel estimation~~
* ~~Modelling of phase discontinuity~~
 |
|  |  |
|  |  |

### Results template for Cases 3/4

Proposal 4a:

For the results template used to collect evaluation results for temporal domain predication 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
	+ CSI scheduling delay
* 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
* Benchmark schemes
	+ Description of benchmark prediction algorithms (e.g., nearest historical CSI and its location, learning window size / time correlation matrix size for auto-regression based prediction, ideal prediction)
	+ Description of feedback schemes, i.e., Rel-18 doppler eT2
* Modelling of spatial consistency
* Modelling of UCI loss (for Case 4; if applicable)
	+ E.g., 10% UCI loss, etc
	+ E.g., 1st UCI is received without error, 10% loss for subsequent UCIs, etc
* Modelling of realistic channel estimation
* Modelling of phase discontinuity

|  |  |
| --- | --- |
| *Support / Can accept* | Futurewei (with comments) |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Futurewei | Same comments as Case 1/2/5, we suggest including model complexity for both the temporal-domain CSI compression model and compared BL model.  |
| Huawei, HiSilicon | 1. Same changes as for Proposal 3a.
2. What is the meaning of “SGCS values before (if applicable) and after compression”? Is it applicable to both AI and benchmark?
3. For the CSI prediction assumptions, 3 aspects are added for report.
4. For “Modelling of UCI loss”, we can add this bullet after how Case 4 works is clarified – if it does not use synchronized CSI information between UE and NW, this item is not needed.
* Temporal domain CSI setting
	+ CSI feedback periodicity
	+ CSI-RS periodicity
	+ ~~CSI scheduling delay~~
* 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
	+ realistic prediction or ideal prediction
	+ separate step or jointly with compression
	+ prediction is AI based or non-AI based
* Benchmark schemes
	+ Description of benchmark prediction algorithms (e.g., nearest historical CSI and its location, learning window size / time correlation matrix size for auto-regression based prediction, ideal prediction)
	+ Description of feedback schemes, i.e., Rel-18 doppler eT2
* Whether/how ~~Modelling of~~ spatial consistency is modeled
* FFS Modelling of UCI loss (for Case 4; if applicable)
	+ E.g., 10% UCI loss, etc
	+ E.g., 1st UCI is received without error, 10% loss for subsequent UCIs, etc
* ~~Modelling of realistic channel estimation~~
* ~~Modelling of phase discontinuity~~
 |
|  |  |

### Results template for multi-vendor training

Proposal 5a:

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 above.

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Lenovo | When discussing multi-vendor cases, we should determine a way to somehow emulate/simulate multiple vendors. Simply having separate dataset from the same EVM does not show multi-vendor case. It is just adding more samples for training. |
| Huawei, HiSilicon | Since Type 2 training collaboration is not adopted in the 5 options of inter-vendor collaboration types in R19, do we still need to consider the corresponding evaluation and results collection in Table 4? |
|  |  |

### Results template for model generalization

Proposal 6a:

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

|  |  |
| --- | --- |
| *Support / Can accept* | Lenovo, Futurewei, New H3C, Huawei, HiSilicon |
| *Object / Have a concern* |  |

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

### Results template for model scalability

Proposal 7a:

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 CSI-RS periodicity

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | For various CSI-RS periodicity, we think it does not impact the model input dimension; it should belong to the generalization case rather than scalability case? |
|  |  |

### Others

Please provide any other comments regarding temporal domain aspects of CSI compression

|  |  |
| --- | --- |
| *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.

**Huawei, HiSilicon:**

***Proposal 2: Avoid duplicated discussion on the modeling of cell/site specific model between CSI compression and CSI prediction, e.g. down selection can be discussed under agenda of CSI compression first.***

**ZTE:**

***Proposal 5: For cell/site specific model, prioritize the alignment on the understandings and EVMs for the cell/site specific model among companies first during Rel-19 study phase.***

* ***Further study and evaluate at least the method of a smaller dataset in the specific cell/site collected for AI/ML model training as a starting point.***

***Proposal 6: For EVM calibration on cell/site specific model, prioritize to construct the dataset for cell/site specific model with the same number of data samples per cell/site compared with generalized model.***

**OPPO:**

***Proposal 7: suggest to study AI/ML based CSI compression with cell-specific 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.***

*Proposal 9: regarding the cell/site/scenario specific model training, two ways can be considered, including*

* ***Direct training based on large cell/site/scenario-specific datasets***
* ***Finetuning based on cell-common model with small cell/site/scenario-specific datasets***

***The trade-off between potential performance gain and complexity/overhead should be further studied.***

*Proposal 3: Regarding the data collection for CSI prediction, cell/site/scenario related “condition information” and “additional condition information” should be considered during the data collection stage.*

***Proposal 4: For the “condition” part, following information should be considered:***

* ***CSI type to be predicted, e.g. raw channel H or eigenvector W,***
* ***CSI-RS configurations, e.g. pattern, time/frequency domain configuration,***
* ***transmission related configuration, e.g. bandwidth and sub-band info, antenna ports, rank, SCS, frequency band,***
* ***cell/site/scenario related information, e.g. Cell ID.***

*Proposal 5: For the “additional condition” part, following information should be considered:*

* ***Cell/site/scenario related information, e.g. region/scenario indication, indoor/outdoor info, UE speed, UE ID, timestamp of data samples, observed SNR***
* ***CSI prediction related information, e.g. observation window, prediction window, sample number/interval.***

**Samsung:**

***Proposal#6: Study the performance of site/cell/location-specific models.***

***Proposal#7: For the evaluation of AI/ML-based CSI prediction using localized models in Release 19, among the options for modeling the spatial correlation in the dataset for a local region consider***

***• A single drop on a single sector with spatial consistency turned on and a large number of UE per drop.***

***Proposal#1: For the evaluation of AI/ML-based CSI prediction using localized models in Release 19, among the options for modeling the spatial correlation in the dataset for a local region consider***

***• A single drop on a single sector with spatial consistency turned on and a large number of UE per drop.***

***Proposal#2: In CSI prediction use case using UE-sided model, consider TRP related aspects for network-side additional condition indication.***

**China Telecom:**

***Proposal 2: Support to study and evaluation the delivery/update of a cell/site-specific model to UE.***

**CMCC:**

*Proposal 1: Option 1 can be considered as baseline 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.***

**LG Electronics:**

**Proposal #3: Prefer option 1 (spatial consistency based) for cell/site specific model evaluation which is optional evaluation.**

**Nvidia:**

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

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

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

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

**Proposal 4: Consider the Madrid grid developed by the METIS project for urban scenarios for further study of site-specific AI/ML based CSI compression.**

**Proposal 5: With a common reference scenario with site specificity, ray tracing is used to generate channel data for the development and evaluation of site-specific AI/ML models for CSI compression.**

**Proposal 2: Site-specific AI/ML models for CSI prediction should be considered to improve performance gain.**

**Proposal 3: Define a common reference scenario with site specificity as a basis for further study of AI/ML based CSI prediction.**

**Proposal 4: Select one the following options to define a common reference scenario with site specificity as a basis for further study of AI/ML based CSI prediction:**

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

**Proposal 5: Consider the Madrid grid developed by the METIS project for urban scenarios for further study of site-specific AI/ML based CSI prediction.**

**Proposal 6: With a common reference scenario with site specificity, ray tracing is used to generate channel data for the development and evaluation of site-specific AI/ML models for CSI prediction.**

**Sony:**

**Proposal 1: To alleviate / resolve some issues related to inter-vendor training, storage and transfer of cell/site-specific AI/ML-based UE-side CSI prediction models, RAN1 will study the following options:**

* **Option 1: Standardized reference model structure + Parameter delivery from NW-side to UE-side**
* **Option 2: Standardized dataset format**
* **Option 3: Standardized model structure + Standardized dataset format**

**Nokia:**

Proposal 4: The proposed scheme of combining a general model with cell/site-specific sub-modules should be considered, regarding the tradeoff between CSI compression performance and model complexity in different cells/sites.

Proposal 5: In addition to cell/site specific models, the use of specific model/feedback overhead within the same cell/site should also be studied.

***Proposal 4:* Consider the following alternative solutions to cope with varying scalability and generalization parameters:**

1. **Scenario specific ML model selection, switching and (de)activation, where each model is trained for specific UE speeds, SINRs, etc.**
2. **One single ML model, or very few ML models, with high generalization and high scalability capabilities.**
3. **UE sided finetuning of generalized ML models based on most recent channel observations over one to few hundreds of ms such that a single ML model can be used in many scenarios.**
4. **Cell and/or location specific retraining of ML models based on training data sets provided by the gNB.**

***Proposal 5:*  Consider in a first step fine tuning performance as an upper bound of what can be achieved by localized models.**

***Proposal 6:* Use CSI-RS burst with K resources and time interval m slots (based on R18 MIMO eType-II) as a starting point for fine tuning methods. Furthermore, consider more efficient triggering of CSI RS configurations for fine tuning.**

***Proposal 7:* We propose to evaluate the benefits of fine tuning as it might be able to adapt to any relevant channel conditions for a single or few generalized ML models. Furthermore, it might provide an upper bound for what can be achieved with cell/cell area specific model retraining without the need to define new localized channel conditions.**

**Qualcomm:**

***Proposal 5: Study techniques and potential specification impact to enable the use of localized models to achieve the associated improvement in the performance-complexity tradeoff.***

***Proposal 6: Capture the following aspects in results table for localized models:***

***• Description of the localized region that is used to train / develop the model***

***• Benchmark scheme contains the model trained under global dataset***

**ITL:**

***Proposal 2: Consider techniques and potential specification impact for the localized models with resolving at least the following issues:***

* ***how to define the local region with specific boundaries with indoor/outdoor states for UEs***
* ***how to maintain the localized model when considering UE mobility***

## Discussions

### LCM aspects

Cell specific models could bring further performance gains on top of cell-common models. The level of this expected gain can be dependent on the evaluation scenario, e.g., payload size, LOS/NLOS ratio etc. On the other hand, there can be additional LCM related discussions including data collection and model management such as model selection, (de)activation, monitoring.

Proposal 11a:

Study LCM aspects of localized models for CSI compression.

* Data collection, training, model pairing, model selection and switching, monitoring
* Any other aspects

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Lenovo | The training procedure and the performance results also should match the scheme proposed for solving the interoperability issue. For example, it is not illustrative to report the performance of localized model trained based on the data of each gNB, and then decide on using Option-1 (specified fixed encoder-or-decoder). |
| Huawei, HiSilicon | Compared with generalized model, there seems no particular spec impact for localized model except for model transfer. Can FL clarify what could be the examples of spec impact for data collection, training, pairing, etc.? |
|  |  |
|  |  |

### Results template

Proposal 12a:

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 with the following additions:

* CSI generation part (localized model)
	+ …
* CSI reconstruction part (localized model)
	+ …
* Dataset description (localized model)
	+ Local region modeling: e.g., Option 1 or Option 2, and further details
	+ Temporal modeling: e.g., how temporal variation is modelled in train and test sets
	+ Train/k (local region)
	+ Test/k (local region)

|  |  |
| --- | --- |
| *Support / Can accept* | Futurewei (with comments) |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Futurewei | For evaluating localized model(s), performance is one aspect and model complexity is another aspect to consider; thus, we suggest including model complexity for both the localized model and the compared BL model, e.g., global model. |
| Huawei, HiSilicon | For “Temporal modeling”, we think we need to align the EVM so that the time gap between training and inference should be at least X sec/min – longer time gap will weaken the time domain correlation between the training samples and testing samples. It is NOT practical for the training side to train and test within a couple of ms/slots.BTW, what is the assumption on the testing data of benchmark? Opt1: An AI solution with generalized model? Opt2: An eType II CB subject to global testing data? Opt3: An eType II CB subject to the same local testing data? If we adopt Opt3, how to ensure the benchmark performance is comparable with the global testing data (which is assumed in R18 evaluation)? |
|  |  |
|  |  |

### Others

Please provide any other comments regarding localized models.

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

# Inter-vendor training collaboration

## Summary of company proposals

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

**Qualcomm Incorporated:**

Table 2: Comparison among 5 options of addressing inter-vendor collaboration complexity

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Inter-vendor collaboration complexity | Performance | Interoperability and RAN4 testing | Feasibility |
| Bilateral collaboration(for baseline comparison) | High | Good | RAN4 may develop a reference CSI reconstruction model | * Feasible for UE implementation
* Forward compatible
 |
| Option 1 | Least effort | * **Without field data**: **Unacceptable**
* **With field data**: **Low to medium** (due to inflexibility of field data used and standardized model structure)
 | Standardized model may be used by RAN4 directly | * Feasible for UE implementation as long as model considers UE capability in terms of model complexity / quantization
* Higher standardization effort
* Not forward compatible
* TBD: feasibility of using field data during standardization
 |
| Option 2 | Least effort | * **Without field data**: **Unacceptable**
* **With field data**: **Low to medium** (due to inflexibility of field data used and standardized model structure)
 | RAN4 may use standardized dataset to develop a reference CSI reconstruction model | * Feasible for UE implementation
* Higher standardization effort
* Not forward compatible
* TBD: feasibility of using field data during standardization
 |
| Option 3 | * **Small** (if UE-side offline engineering allowed, i.e., **2a** above)
* **Larger** (if UE inference using transferred model, i.e., **2b** above)
 | **Low to medium** (due to inflexibility of standardized model structure) | RAN4 may develop a reference CSI reconstruction model | * If UE-side offline engineering allowed **(2a)**: **feasible for UE implementation**
* If UE inference using transferred model **(2b): feasible if and only if parameters are fully tested at the UE-side prior to exchange**
* Otherwise, Not feasible for UE implementation
* May not be forward compatible
 |
| Option 4 | Small  | Good | RAN4 may develop a reference CSI reconstruction model | * Feasible for UE implementation
* Forward compatible
 |
| Option 5 | * **Small** (if UE-side offline engineering allowed, i.e., 2a above)

**Larger** (if UE inference using transferred model, i.e., 2b above) | Good | RAN4 may develop a reference CSI reconstruction model | * If UE-side offline engineering allowed (2a): feasible for UE implementation
* If UE inference using transferred model (2b): feasible if and only if parameters are fully tested at the UE-side prior to exchange
* Otherwise, Not feasible for UE implementation
* Forward compatible
 |

1. RAN1 should further study how to include real world data in option 1 and option 2.
2. To address concerns related to inter-vendor training collaboration complexity in option 3 / 4 / 5, study standardization support of registering/retrieving parameters / dataset / reference model from a central registry.

**Huawei, HiSilicon:**

***Proposal 7: Capture Table 8 to the TR 38.843 for the comparison over the options of inter-vendor training collaboration.***

Table 8 Comparison over the 5 options

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Inter-vendor collaboration complexity** | **Performance** | **RAN4 / testing related aspects** | **Feasibility** |
| Option 1 | Relieved  | Severely Limited | Testable | Less challenging for realization. Large spec effort/spec evolution |
| Option 2 | Relieved | Severely Limited | Testable | Less challenging for realization. Large spec effort/spec evolution |
| Option 3 | Relieved if the parameter exchange is performed with standardized procedure | Partially Limited | FFS | Depends on feasibility of model transfer/delivery for realization. Non-trivial spec effort/spec evolution but less than Option 1/2 |
| Option 4 | Relieved if the dataset exchange is performed with standardized procedure | Less Limited | FFS | Depends on whether dataset delivery is achieved with low collaboration complexity. Less spec effort than Option 1/2/3. |
| Option 5 | Not relieved (under Case y/z2) | Less Limited | FFS | Depends on feasibility of model transfer/delivery for realization. Spec effort depends on whether/how to specify model representation format. |

**Futurewei:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Collaboration complexity** | **Performance** | **Interoperability and testing aspects** | **Feasibility** |
| **Option 1**(Fully standardize reference model) | Least complexity among the 5 options | May be impacted compared to vendor (UE or NW) specific model(s).  | Least testing effort involved among the 5 options. | More discussion time and effort required to agree on the reference model structure and parameters. |
| **Option 2**(Standardize dataset) | Significant collaboration effort may still be involved.May sub-categorize into:2a) Reference model structure (including quantization part) is specified.2b) Reference model structure is not specified. | Model performance depends on vendor implementation and potential optimization of the encoder and decoder. | Significant effort may still be involved. | In addition to issues related to pairing/alignment between encoder and decoder and between quantization and dequantization methods and codebook(s), standardizing dataset(s) may have privacy concerns. |
| **Option 3**(Standardized reference model structure + Parameter exchange between NW-side and UE-side) | Depending on whether the dataset is standardized and what training collaboration type is used, collaboration complexity may be different.May sub-categorize into:3a) Dataset is specified.3b) Dataset is not specified. | Compared to vendor (UE or NW) specific model structure(s), performance may be impacted. Given that parameters are based on vendor implementation, some level of optimization can be achieved by vendor. | Depending on whether the dataset is standardized and what training type is used, significant interoperability testing effort may still be involved in the case that vendors choose to use different datasets or dataset formats. | This option requires less time in agreeing reference model structure only among companies, compared to Option 1. |
| **Option 4**(Standardized data / dataset format + Dataset exchange between NW-side and UE-side) | Significant collaboration effort may still be involved.May sub-categorize into:2a) Reference model structure is specified.2b) Reference model structure is not specified. | Performance would be very similar to what was studied in Rel-18 if reference model structure is not specified (model is completely dependent on vendor implementation). | This option can reduce some collaboration effort compared to the case in which each vendor may choose its own data/dataset format, however, major issues remain to be resolved. | This option may require less effort in reaching consensus (more feasible than options 1 – 3). |
| **Option 5**(Standardized model format + Reference model exchange between NW-side and UE-side) | If only model format is specified, significant collaboration effort is still needed. | The performance for using this option alone would be very similar to what was studied in Rel-18 (model is completely dependent on vendor implementation). | This option can reduce some collaboration effort compared to the case in which each vendor may choose its own data/dataset format, however, major issues remain to be resolved. | This option may require less effort in reaching consensus (more feasible than options 1 – 3). |

***Proposal 1: For CSI compression using two-sided model sub use case, consider further studying the common dataset format (Option 4) and/or common model format (Option 5) as a starting point for alleviating inter-vendor collaboration effort.***

***FFS: what data/dataset format and model format to use.***

**Spreadtrum Communications:**

***Proposal 2: For options to alleviate / resolve the issues related to inter-vendor training collaboration of two-side model.***

* ***For option 1, option 3 and option 5, we can further discuss them until good progress from RAN4 perspective.***
* ***For option 2, it will require great effort to ensure multiple models can be trained.***
* ***For option 4, definition of data/dataset format should be discussed first.***

**Intel Corporation:**

***Proposal 6***:

* *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.*

|  |  |  |
| --- | --- | --- |
|  | **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 3** | CSI generation part weights are transferred/delivered from NW side to UE side: **Type 1 NW side training,** CSI generation part structure cannot change.CSI reconstruction part weights are transferred/delivered from UE side to NW side: **Type 1 UE side training,** CSI reconstruction part structure cannot change. |
| **Option 4** | Dataset is transferred/delivered from NW side to UE side: **Type 3 NW-first**.Dataset is transferred/delivered from UE side to NW side: **Type 3 UE-first**. |
| **Option 5** | CSI generation part model is transferred/delivered from NW side to UE side: **Type 1 NW side training**.CSI generation part model is transferred/delivered from NW side to UE side: **Type 1 UE side training**. |

***Proposal 7***:

* *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 8***:

* *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 typeCharacteristics | 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 9***:

* *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 typesCharacteristics | 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 |

**vivo:**

**For the study of inter-vendor collaboration issue, consider the timescale of deployment when comparing different options.**

**Characteristics of options to alleviate / resolve the issues related to inter-vendor training collaboration of AI/ML-based CSI compression can be summarized as:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Inter-vendor collaboration complexity** | **Performance** | **Interoperability and RAN4 / testing related aspects** | **Feasibility** | **Deployment timescale** |
| **Option 1** | Minimum complexity | Restricted | Solved | feasible | \ |
| **Option 2** | Minimum complexity | Highly restricted | Not solved | feasible | \ |
| **Option 3** | Minimum complexity with over the air signalling; Otherwise high; | Optimum  | Solved | Feasible | Short |
| **Option 4** | High complexity in server to server manner; Medium complexity with over the air signalling;  | Better than Option1/2, but worse than Option3 and Option5 | Not solved | Infeasible with only on-device operation | Long |
| **Option 5** | High complexity in server to server manner; Medium complexity with over the air signalling;  | Optimum | Not solved | Infeasible with only on-device operation | Long |

**Option 3 should be supported, and then further study how to standardize reference model.**

**Option 3 should be supported, and then further study how to standardize reference model.**

**The reference model structure may be aligned through the following procedures**

* + - * **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.**

**ZTE:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Inter-vendor training collaboration options | Inter-vendor collaboration complexity | Performance | Interoperability and RAN4 / testing related aspects | Feasibility |
| Option 1 | Not need inter-vendor collaboration  | Sub-optimal, limited to the standardized reference model  | Interoperable and compatible to RAN4 testing | Feasible* Much specification workload
* Specification evolution with the development of AI model
 |
| Option 2 | Not need inter-vendor collaboration | Sub-optimal, limited to the standardized dataset | Interoperable | Feasible* Much specification workload
* Multiple sets of datasets specified for multiple scenarios
* Specification evolution with the development of AI model
 |
| Option 3 | Need parameter exchange between NW-side and UE-side, * Over-the-air parameter transfer may result in some resource overhead, less complexity compared with Option 4 and Option 5
* Offline parameter delivery may result in additional offline multi-vendor collaboration
 | Sub-optimal, limited to the specified reference model structure, however, better than Option 1&2 | Interoperable if parameter type is aligned between NW side and UE side | Feasible * Delivered model parameters need to be interpreted at the other side
 |
| Option 4 | Need dataset exchange between NW-side and UE-side* Over-the-air dataset delivery may result in huge resource overhead
* Offline dataset delivery may result in additional offline multi-vendor collaboration
 | Sub-optimal, performance loss compared with Option 5 according to TR 38.843 | May be interoperable, however, performance loss may incur due to the misaligned model backbone according to TR 38.843  | Less feasible* Over-the-air dataset delivery may result in huge resource overhead, large UE power consumption, large latency
 |
| Option 5 | Need reference model exchange between NW-side and UE-side, less complexity than Option 4* Over-the-air model transfer may result in large resource overhead
* Offline model delivery may result in additional offline multi-vendor collaboration
 | Optimal compared with other options | May not be interoperable due to the misaligned hardware/software capability at the other side  | May not be feasible* UE/NW may not operate the exchanged unknown reference model due to misaligned hardware/software capability
 |

***Proposal 7:*** *Deprioritize the discussion on the approaches beyond RAN1 scope in this agenda item, which should be left to be discussed in other working groups.*

***Proposal 8:*** *To alleviate / resolve the issues related to inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model, Option 1 (Fully standardized UE-part reference model) and Option 3 (Model transfer z4) can be studied as a starting point.*

***Proposal 9****: For comprehensive analysis on AI/ML framework for two-sided model, further study a complete and unified solution for model identification, multi-vendor collaboration, and model pairing.*

**Google:**

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

**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: prefer Option 3 and Option 4 with fewer standardization work and implementation flexibility.*

**CATT:**

Proposal 7: 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 8: 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.

**Samsung:**

|  |  |  |
| --- | --- | --- |
| **Training collaboration type** |  **Potential Advantages**  | **Requirements and Limitations**  |
| **Option 1**: Fully standardized reference model (structure + parameters) | * No requirement for multi-vendor offline collaboration
* Guaranteed inter-operability
* Suitable for RAN4 test
* Vendors can develop their own proprietary model compatible to the reference model (for hardware optimization, complexity reduction, etc.)
 | * Performance fixed to the standardized reference model, e.g., inability to adapt to sites/ non-considered scenarios/configurations.
* High standardization effort (difficulty to reach to one agreeable model).
 |
| **Option 2*:*** Standardized dataset | * No requirement for multi-vendor offline collaboration
* Guaranteed inter-operability
* Suitable for RAN4 test
* Vendors can develop their own proprietary model compatible to the standardized dataset (for HW optimization, complexity reduction, etc.)
 | * Performance fixed to the standardized reference model, e.g., inability to adapt to sites/ non-considered scenarios/configurations.
* High standardization effort (difficulty to reach to one agreeable dataset).
 |
| ***Option 3:*** Standardized reference model structure + Parameter exchange between NW-side and UE-side | * No requirement for multi-vendor offline collaboration
* Better performance than Option 1 and Option 2 (adaptation to sites/cells)
* Guaranteed inter-operability
* Some difficulties for RAN4 test ( model parameter could be per site/cell)
* Model transferring vendors can develop their own proprietary models
 | * Capability for model transfer (receive a parameter and run in plug-and-play manner)
* Restriction on model structure to the standardized structures.
* Performance mainly controlled by the transferring vendor, e.g., NW vendor, if NW to UE.
* High standardization effort (difficulty to reach to a set of agreeable structures).
 |
| ***Option 4:*** Standardized data / dataset format + Dataset exchange between NW-side and UE-side | * Potentially better performance than Option 1 and Option 2 (model adaptation to sites/cells, if dataset is shared per sites/cells)
* Guaranteed inter-operability
* Difficulties for RAN4 test ( dataset could be per sites/cells)
* Vendors can develop their own proprietary models
* Lower standardization (dataset format and exchange mechanism) effort compared to Option 1-3
 | * Performance mainly controlled by the dataset transferring vendor, e.g., NW vendor, if NW to UE.
* Slower/longer dataset exchange to model deployment cycle, i.e., does not work in plug-and-play manner (receiving vendor collects the dataset and train its model offline)
* If the dataset exchange is not peer-to-peer, e.g., gNB and UE, absence of central entity
* Standardization effort for dataset format and exchange mechanism
 |
| ***Option 5:*** Standardized model format + Reference model exchange between NW-side and UE-side | * Potentially better performance than Option 1 and Option 2 (model adaptation to sites/cells, if reference model is shared per site/cell)
* Guaranteed inter-operability
* Difficulties for RAN4 test ( reference models could be per site/cell)
* Vendors can develop their own proprietary models compatible to the shared reference model.
* Lower overhead than Option 5, if the models are exchanged OTA.
* Lower standardization (dataset format and exchange mechanism) effort compared to Option 1-3
 | * Performance mainly controlled by the dataset transferring vendor, e.g., NW vendor if NW to UE.
* Slower/longer dataset exchange to model deployment cycle, i.e., does not work in plug-and-play manner (receiving vendor collects the dataset and train its model offline)
* Standardization effort for dataset format and exchange mechanism
 |

**The following guiding principles can be considered:**

1. **UE-side vendor-agnostic AI/ML model/solution at the network**
2. **UE-side AI/ML solutions shall consider different classes of UEs**
3. **Efficiency and convenience**

**Ericsson:**

Align the proposals for standardized reference/test encoder/decoder (parameters) between RAN1 and RAN4, and conclude that the feasibility study of this option is handled by RAN4.

Studying the feasibility of RAN1 Option 1 should be prioritized. RAN1 should down prioritize the study of Option 2, 3, and 4 until further progress being made regarding the feasibility of RAN1 Option 1 (RAN4 Option 3 or 4).

**China Telecom:**

**BJTU:**

**CMCC:**

*Proposal 3: To alleviate / resolve the issues related to inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model, suggest focusing on the option 1, option 3 and option 5.*

*Proposal 4: Regarding option 1 to alleviate / resolve the issues related to inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model, suggest leaving the discussion on details of standardized model to RAN4.*

*Proposal 5: Regarding option 5 to alleviate / resolve the issues related to inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model, suggest discussing the mechanism of model transfer/delivery firstly.*

**LG Electronics:**

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

**Proposal #5. Study on model complexity method, e.g., knowledge distillation, to further reduce the CSI training/signaling complexity for Type 3 training collaboration.**

**Xiaomi:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  ***Related aspects******Options*** | ***Inter-vendor collaboration complexity*** | ***Performance*** | ***Interoperability and RAN4 / testing*** | ***Feasibility*** |
| ***Option 1*** | ***Less*** | ***Limited*** | ***Recommended to discuss in RAN4*** | ***Yes*** |
| ***Option 2*** | ***Less*** | ***Limited and better than Option 1*** |
| ***Option 3*** | ***Depend on how to deliver model parameter*** | ***Better than Option 1***  |
| ***Option 4*** | ***Depend on how to deliver dataset***  | ***Better than Option 3*** |
| ***Option 5*** | ***Depend on how to deliver reference model*** | ***Similar with Option 4*** |

**Panasonic:**

**Proposal 1: For Option 2, two candidates to realize the standardized dataset can be considered.**

* **Option 2-1: 3GPP specific dataset is specified.**
* **Option 2-2: 3GPP dataset database is specified.**

**Proposal 2: Option 5 is deprioritized in Rel.19 study.**

**Observation 13:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Inter-vendor collaboration** | **Performance** | **Interoperability / RAN4 testing** | **Feasibility** |
| Option 1 | Not required | Depends on standardized reference model | Less IOT testing effort is required. | Feasible |
| Option 2-1:3GPP specific dataset is specified. | Not required | Depend on the standardized dataset. | Less IOT testing effort is required (but depending on the number of specified datasets). | Feasible |
| Option 2-2:3GPP dataset database is specified. | Not required | The datasets are the result of the specific environment / implementation (of gNB, UE, site, channel conditions, etc.) | IOT testing effort depends on the amount of combination of datasets. | Feasible |
| Option 3 | Parameter exchanges | Potential to support localized model by updating the parameters | How to maintain interoperability should be considered. | FFS |
| Option 4 | Dataset exchange | Potential to support localized model by exchanging the dataset specific to localized model. | How to maintain interoperability should be considered. | FFS |
| Option 5 | Reference model exchange | Potential to support localized model by exchanging the reference model specific to localized area. | How to maintain interoperability should be considered. | FFS |

**NEC:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  Option typeCharacteristics | Option 1 | Option 2 | Option 3 | Option 4 | Option 5 |
| Inter-vendor collaboration complexity | Little | 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 3: 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 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***

***Proposal 4: The privacy/proprietary should be considered for the comparison between dataset delivery (Case 4) and model delivery (Case 3/5).***

**Fujitsu:**

***Proposal 3:***

* *RAN1 further study how to use the standardized reference model for alleviating / resolving the issues related to inter-vendor training collaboration for AI/ML-based CSI compression using two-sided model. Specifically, two sub-options could be added for Option 1 and Option 3, respectively:*
	+ *Option 1: Fully standardized reference model (structure + parameters)*
		- *Option 1-1: The fully standardized reference model is used in UE/NW for generating/reconstructing the CSI.*
		- *Option 1-2: The fully standardized reference model is used as a reference to train the CSI generation/reconstruction models.*
	+ *Option 3: Standardized reference model structure + Parameter exchange between NW-side and UE-side*
		- *Option 3-1: The standardized reference model structure is used in UE/NW for generating/reconstructing the CSI.*
		- *Option 3-2: The standardized reference model structure is used as a reference to train the CSI generation/reconstruction models.*

***Proposal 4:***

* *RAN1 to study the feasibility of using fully standardized reference models for CSI feedback, from at least the following aspects:*
	+ *Proprietary issue.*
	+ *Performance superiority.*
	+ *Hardware optimization.*

***Proposal 5:***

* *The dataset of ground-truth CSI for training should be standardized if standardized reference models are used as a reference to train the actual AI/ML models for performing the CSI feedback.*

***Proposal 6:***

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

***Proposal 7:***

* *If fully standardized reference models are used, either as a reference to train the actual AI/ML models or directly used for CSI feedback, which part(s) of the reference model to be standardized should be studied.*
	+ *Alt 1: standardized reference CSI generation model.*
	+ *Alt 2: standardized reference CSI reconstruction model.*
	+ *Alt 3: standardized pair of reference CSI generation and reconstruction models.*

***Proposal 8:***

* *To alleviate/resolve the issues related to inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model using Option 3, the method of generating and delivering the parameters should be clarified, including at least:*
	+ *How the parameters are determined, e.g., by a specific vendor, or vendor collaboration?*
	+ *How to share the parameters among UE vendors or among the NW vendors, if needed?*

***Proposal 9:***

* *To alleviate/resolve the issues related to inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model using standardized dataset, RAN1 to further study the methods of generating the dataset, considering the following aspects as a starting point:*
	+ *The size of the dataset.*
	+ *The scenarios that the dataset cover.*
	+ *The format of the data/dataset.*

***Proposal 10:***

* *To alleviate/resolve the issues related to inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model using Option 4, the method of generating and delivering the dataset should be studied, including at least*
	+ *How the dataset is determined, e.g., by a specific vendor, or vendor collaboration?*
	+ *How to share the dataset among UE vendors or among the NW vendors, if needed?*

***Proposal 11:***

* *To alleviate/resolve the issues related to inter-vendor training collaboration of AI/ML-based CSI compression using two-sided model, RAN1 to further study the feasibility and complexity of Option 5.*

**InterDigital, Inc.:**

**NVIDIA:**

**Apple:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Inter-vendor collaboration complexity**  | **Performance** | **Interoperability and testing** | **Feasibility**  |
| **Option 1: Std ref model** | **Low** | **Limited. The standard ref model is performance upper bound** | **Yes** | **Hard to define a ref model is specification** |
| **Option 2: Std dataset** | **Low** | **Limited, if dataset does not match real environment** | **Yes** | **Hard to define a dataset is specification** |
| **Option 3: Std ref model structure + param exchange** | **High** | **Good.** | **Challenging for RAN4 test** | **Challenging, but easier than full model.** |
| **Option 4: Std data format + dataset exchange** | **Medium** | **Good.** | **Challenging for RAN4 test** | **Yes** |
| **Option 5: Std model format + ref model exchange** | **High if model structure is offline aligned** | **Good** | **Challenging for RAN4 test** | **Limitation identified in model transfer z4 and z5** |

**Lenovo:**

Considering “Standardized reference model or dataset” methods for resolving inter-vendor training collaboration complexity, capture the following benefit and challenges:

|  |  |
| --- | --- |
|  | Standardized reference model or dataset |
| Inter-vendor collaboration | Very effective - remove the need for bilateral collaboration by providing a reference model for design of the encoder/decoder |
| Performance | Limiting - reduce the achievable performance due to lack of possibility of capturing the statistics of the data for different UE-NW/scenarios/conditions |
| Interoperability, RAN4 Testing | Very effective - provides a reference model facilitating RAN4 performance requirements |
| Feasibility | Not clear Feasibility of designing such standardized reference model(s) with good performance in difference scenarios (different vendor pairs) is not clear.Feasibility of collecting good training data for training of such model should be investigated.  |

Considering “common framework for exchange of model/dataset/weights” methods when the NW exchange information regarding the “local encoder” for resolving inter-vendor training collaboration complexity, capture the following benefit and challenges:

|  |  |
| --- | --- |
|  | Common Framework – Lcoal Encoder model exchaneg |
| Inter-vendor collaboration | Very effective for NW sideEffective with some complexity for UE side: the UE side still may need to maintain different models for different NW node |
| Performance | Improved performance compared to “standardized encoder/decoder” format as different NW nodes can have their model adapted to the real-world data.May have very low performance for UEs (UE types) which were not present during training data collection. |
| Interoperability, RAN4 Testing | Possible to handle - If RAN4 defines a test encoder/decoder, it can be used for training of an Encoder at the UE and also can be considered during the training of the two-sided model at the NW side.  |
| Feasibility | Feasible |

Considering “common framework for exchange of model/dataset/weights” methods when the NW exchange information regarding the “local encoder” for resolving inter-vendor training collaboration complexity, capture the following benefit and challenges:

|  |  |
| --- | --- |
|  | Common Framework – DEcoder model exchaneg  |
| Inter-vendor collaboration | Very effective for NW sideEffective with some complexity for UE side: the UE side still may need to maintain different models for different NW node also need to train the encoder model for each NW side. |
| Performance | Best performance between different options, i.e., different NW nodes can have their model adapted to the real-world data and also new UEs (UE-types) can adapt the encoder based on their input data statistics.  |
| Interoperability, RAN4 Testing | Possible to handle - If RAN4 defines a test encoder/decoder, it can be used for training of an Encoder at the UE and also can be considered during the training of the two-sided model.  |
| Feasibility | Feasible |

**Sony:**

|  |  |  |
| --- | --- | --- |
|  | Pros | Cons |
| Option 1 | * Signaling overhead would be less than other options.
* gNB and UE don’t need to train at each side.
 | * Spec impact would be higher than other options.
* Low AI/ML model performance differentiation between vendors.
 |
| Option 2 | * High AI/ML model performance differentiation between vendors.
* Signaling overhead would be lower than option 4.
 | * Two-sided model incompatibility could occur .
* Spec impact could be higher than option 4.
 |
| Option 3 | * Two-side model compatibility can be assured.
* Spec impact could be lower than option 1.
 | * Signaling overhead would be larger than option 1 and 2.
 |
| Option 4 | * High AI/ML model performance differentiation between vendors.
* Spec impact could be lower than option 2.
 | * Two-side model incompatibility could occur.
* Signaling overhead could be larger than option 2.
 |
| Option 5 | * Two-side model compatibility can be assured.
* Spec impact could be lower than option 1 and 3.
 | * Signaling overhead would be larger than option 1, 2 and 3.
 |

**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 model.**

**Nokia:**

**Proposal 6:** **For support of interoperability in CSI compression using a two-sided model, RAN1 shall consider standardization of the reference CSI reconstruction part (CSI decoder) model (Option 3 without trained parameter exchange between NW-side and UE-side). As regards model training collaboration types, RAN1 shall focus on NW-first approaches, i.e., Type 2 Sequential (via API sharing) and Type 3 NW-first (via training dataset sharing) to facilitate UE-side model training.**

**Proposal 7: As regards CSI compression using a two-sided model, RAN1 shall agree on definition of the reference model structure in technical terms. One important attribute of the reference model, to our thinking, is as below. 3GPP needs to align on additional attributes and/or requirements for the reference model, if any.**

* **A reference model is a model which can be used for two-sided model training, and re-training is not required once the model training procedure is once completed using a reference model, even though the actual model in use can be not exactly same as the reference model.**

**Proposal 8:** **Regarding standardization of the reference CSI reconstruction part model structure for CSI compression using a two-sided model use case, RAN1 shall agree on definition of the input and output interface/data format of the reference CSI reconstruction part model.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Inter-vendor collaboration complexity | Main concern(s) | Verdict | Additional benefit |
| Option 1 | Lowest | E2E performance concernNo or small room for possible future enhancements | Not recommended. | Standardization of CSI reconstruction part: can be re-used for RAN4 test decoder definition |
| Option 2 | High | Practical feasibility questionableNo big benefit expected in terms of inter-operability complexity reduction as a stand-alone option  | Not recommended. |  |
| Option 3(- parameter exchange part)(+ Option 4 aspects of training dataset sharing) | Manageable | Agreement on definition of reference model structure required | Recommended.Standardization of reference decoder model structure in conjunction with NW-first training scheme, possibly w/standardization of input/output data format of the reference decoder  | Proxy-DEC model or Intermediate KPI predictor-based model monitoring easily supportableStandardization of reference CSI reconstruction part model structure: can alleviate efforts in RAN4 test decoder definition |
| Option 4 | High | No big benefit expected in terms of inter-operability complexity reduction as a stand-alone option | Not recommended as stand-alone.Better to be combined with Option 3 |  |
| Option 5 | Needs to be clarified further.Similar pros/cons expected to Option 3, with more degree of freedom in proprietary model selection at UE-/NW-vendors at the cost of higher inter-vendor collaboration complexity |

Proposal 9: 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.

**ETRI:**

**Observation 1: For Options based on exchanging the reference model (i.e., Options 1, 3, and 5), performance degradation of non-reference models compared to the reference model can exist.**

**Observation 2: For Options based on the sharing of fully standardized models or datasets (i.e., Options 1 and 2), large standardization efforts are expected to be required.**

**Observation 3: For Options based on the sharing of fully standardized models or datasets (i.e., Options 1 and 2), assessments of generalized performances are important, because the number of standardized models or datasets may be limited due to the large standardization efforts.**

**Observation 4: When considering over-the-air-interface delivery, inter-vendor training based on the dataset delivery (i.e., Option 4) is less feasible, due to the size of datasets and additional training time.**

**Observation 5: When considering over-the-air-interface delivery, inter-vendor training based on the reference model delivery (i.e., Options 3 and 5) is less feasible, due to the additional training time. Conversely, delivery of the reference model (i.e., Options 3 and 5) for direct use is feasible when it is applicable.**

**Observation 6: When considering offline delivery, inter-vendor training based on the datasets or reference model delivery (i.e., Options 3~5) is feasible.**

**CEWiT:**

**Proposal-4: 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-5: Consider the possibility of using model-ID based standardized model for the same architecture with different configurations.**

**Proposal-6: Consider a common pre-processing technique to ensure standardized dataset.**

**Proposal-7: The model parameter exchange between the UE side and the NW side should be specified after the model identification and selection process.**

**Proposal-8: Consider using model ID based identification for ensuring proper training between UE sided model and NW sided model**

**Proposal-9: 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-10: 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-11: In case of Type-III UE first raining, train the CSI reconstruction model with the knowledge of UE specific codebook.**

**MediaTek:**

If downselection is needed, prioritize UE-first sequential separate training.

Prioritize standardization of at least a CSI generation part at structure level.

**SK Telecom:**

**CAICT:**

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Inter-vendor collaboration complexity** | **Performance** | **Interoperability** | **Feasibility** |
| **Option 1** | **Low** | **Limited** | **Easy to implement** | **Feasible and high specification workload requires** |
| **Option 2** | **Low** | **Limited** | **Easy to implement** | **Feasible and high specification workload requires** |
| **Option 3** | **Medium** | **Good** | **Standardized dataset is required to ensure interoperability** | **Feasible and model transfer should be supported** |
| **Option 4** | **High** | **Good** | **Standardized dataset is required to ensure interoperability** | **FFS** |
| **Option 5** | **Medium** | **Good** | **FFS** | **FFS** |

**NTT DOCOMO, INC.:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Option 1 | Option 2 | Option 3 | Option 4 | Option 5 |
| **Collab. complexity** | Low | Low | Medium | High | Medium |
| **Performance** | Less flexibility in using the model with the best performance. | Same as Option 1 with additional performance loss caused by different training schemes.  | Less flexibility in using better model structures.  | Performance loss by the training schemes, especially with different backbone structures. | Depending on implementation. |
| **Interoperability/RAN4 aspects** | Interoperable and testable. | Interoperable. Additional schemes may be required for RAN4 aspects. | FFS. At least less interoperable compared to Option1/2. | FFS. At least less interoperable compared to Option1/2. | Difficult for interoperation and RAN4 aspects. |
| **Feasibility** | Feasible. | Feasible. | FFS based on detailed schemes. | FFS based on detailed schemes. | Infeasible. |

**Proposal 2**

* **Study the methods for performance improvement with Option 1/2.**
* **Study the detailed schemes to identify interoperability and feasibility for Options 3 and 4.**
* **Deprioritize Option 5 during Rel. 19 study.**

**ITL:**

***Proposal 3: Both option 1 and option 3 are slightly preferred since these options facilitate collaboration and compatibility among vendors and provide clarity and consistency through the standardization process without the issues related to ownership and proprietary rights***

**Fraunhofer IIS, Fraunhofer HHI:**

**IIT Kanpur, Indian Institute of Tech (M):**

**Proposal 1: Include a hybrid option 6 for study:**

**Option 6: Standardized reference model structure + Standardized reference dataset**

**Indian Institute of Tech (M), IIT Kanpur:**

## Discussion

|  |
| --- |
| **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 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.***
 |

Below is a quick summary of analysis provided by companies.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Inter-vendor collaboration complexity** | **Performance** | **Interoperability and RAN4 testing** | **Feasibility** |
| **Bilateral collaboration****(for baseline comparison)** | High | Best |  |  |
| **Option 1** | Low / None  | Limited  | GoodRelated to RAN4-Option3 and RAN4-Option4 | Feasible for UE implementation.Large spec effort (initially & for evolution) |
| **Option 2** | Low / None  | Limited | Good, though some companies disagree.Related to RAN4-Option4 | Feasible for UE implementation.Large spec effort (initially & for evolution) |
| **Option 3** | Mixed assessments (low, manageable, medium, high) that may depend on sub-flavours | Somewhat limited / Best | Mixed assessments (need RAN4 study, FFS, possible, solved, challenging, poor) | Feasible for UE implementation w/ UE-side offline engineering.Feasibility for UE implementation for plug-and-play inference (Case z4) depends on UE capability.Large spec effort (initially & for evolution) but less than Options 1/2 |
| **Option 4** | Mixed assessments (low, medium, high) that may depend on sub-flavours | Best / somewhat limited | Mixed assessments (need RAN4 study, questionable, challenging, poor, possible, interoperable) | Feasible for UE implementation w/ UE-side offline engineering.Less spec effort. |
| **Option 5** | Mixed assessments (low, medium, high) that may depend on sub-flavours | Best | Mixed assessments (need RAN4 study, questionable, challenging, poor, possible, interoperable) | Feasible for UE implementation w/ UE-side offline engineering.Feasibility for UE implementation for plug-and-play inference of known structure (Case z4) depends on UE capability.Feasibility for UE implementation for plug-and-play inference of unknown structure (Case z5) is questionable.Less spec effort. |

Main observations

* Option 2 shares similar characteristics as Option 1. In fact, developing a standardized dataset in Option 2 requires developing reference models during the standardization process. In that sense, Option 2 shares the same challenges (large standardization effort, limited performance) as Option 1, but Option 2 may have some disadvantages compared to Option 1, and the benefit of Option 2 over Option 1 is unclear. So, it seems safe for RAN1 to deprioritize Option 2 in view of Option 1.
* Option 1/2 has good interoperability and RAN4 testability, so it will be good to have Option 1/2 supported (either through RAN4, RAN1, or both), so that interoperability and RAN4 testability questions are resolved. It’s also noted by companies that RAN1’s Option 1 resembles either RAN4-Option3 or some potential flavor of RAN4-Option4 and that RAN1’s Option 2 resembles some potential flavor of RAN4-Option4.
* Quite a few companies pointed out performance limitations with Option 1/2 due to generalization capability concerns, limitations of specified model(s), difficulty of realizing site-specific models, etc. So, it seems obvious that Option 1/2 alone will not be enough, and that we will have to consider one or more of Option 3/4/5.
* In terms of performance, Option 5 is least restricted and can deliver the best performance. Options 3 and 4 may have some performance limitations due to fixed model structure and/or potential loss due to suboptimal training but such performance limitation should be minor and do not affect feasibility.
* The feasibility of Option 3/5 for UE implementation depends on:
	+ (1) Expectation at the UE / UE-side with the exchanged model/parameters. That is, whether the model/parameters are for immediate use for inference (i.e., plug-and-play inference) or whether UE-side is expected/allowed to do further UE-side offline engineering to re-develop, retrain, quantize, and test a model. Note that the UE-side offline engineering is one-sided and does not require inter-vendor collaboration.
	+ (2) UE capability, in case the model/parameters are for immediate use for inference
	+ So, we may have to discuss those two cases separately in further discussion.
* Inter-vendor collaboration complexity may also depend on the expectation at the UE / UE-side (i.e., plug-and-play inference or offline further engineering), as well as the methods of exchanging the parameters / dataset / reference model.

In general, I think the pros/cons analysis of Option 1-5 are clear and companies’ views and understandings are largely aligned. So, we should not repeat/debate the same discussion in the next meeting, and it may not be productive to try to capture/agree the pros/cons table. Rather, we should focus on moving the discussion to the next step in a constructive manner.

Toward such a goal, here are some initial thoughts on what I think the group can try to achieve and how the group can move forward. I hope that every company is on board along this direction.

* **Specify Option 1 (either by RAN4 or by both RAN4 and RAN1)**, at least for interoperability and RAN4 testing
	+ RAN4 is already discussing RAN4-Options3 and RAN4-Option4, so RAN1 does not need to repeat RAN4’s work.
	+ RAN1 can discuss whether specified model(s) from Option 1 is only for interoperability and RAN4 testing, in which case it can be entirely handled by RAN4, or if the specified model(s) are also for use in the field. If former, the Option 1 specification can be handled by RAN4. If latter, RAN1 may want to discuss whether RAN1 wants to specify additional models.
* **Also, specify one or more Options among Option 3/4/5** to allow either parameter update on standardized model(s) or use more powerful proprietary model structures.
	+ Option 3 requires specifying model structures, so RAN1 may start discussions on aspects related to model structures with Option 3 in mind. I don’t think RAN1 needs to discuss detailed model structures at this point, but rather discuss aspects related to the model structure specification, such as input/output, quantization, generalization/scalability, complexity, etc. to assess feasibility.
		- The model structure discussions can obviously be leveraged toward Option 1 if needed.
	+ Take device feasibility into account (i.e., whether plug-and-play is possible/desirable) in further discussions. In the end, we may consider potentially two different solutions for the following two different types of scenarios that may relate to device capability, degree of hardware optimization that a vendor wants to do, and/or engineering practices:
		- Model/parameters received at the device are directly used for inference at the device in a plug-and-play manner (assuming such device capability)
		- Model/parameters/dataset received at the device are used for offline engineering at the device-side vendor to potentially develop and optimize a different model/parameters used for inference at the given device and other devices from the vendor.
	+ Study ways to alleviate inter-vendor interoperability concerns for Option 3/4/5. Let’s focus on solutions that make sense in resolving inter-vendor collaboration complexity, are scalable, and are feasible for UE implementation. We may want to discuss various methods of exchanging the parameters/dataset/model.
		- Some companies mention that certain flavours are outside RAN1 scope, but I respectfully disagree. Without RAN1 first identifying a potential method, other working groups will not even discuss it. Therefore, it’s RAN1’s role to identify potential methods and let/ask other working groups look at whether/how the methods work.
		- To simplify the discussion, I think we can focus on the model/parameter/dataset exchange originating from the NW-side and ending at the UE-side.
			* However, this does not mean that we’re deprioritizing the exchange originating from the UE-side and ending at the NW-side from potential specification.
	+ Some of the analysis for Option 3/5 depends on whether the specified/exchanged model is the CSI generation part or reconstruction part, or both. So, it will be good to have discussions on these sub-flavours.

Proposal 21a:

* Specify, if feasible, Option 1, at least for interoperability and RAN4 testing
	+ RAN1 to further discuss whether specified model(s) from Option 1 is only for interoperability and RAN4 testing, in which case it can be handled by RAN4, or if the specified model(s) are also for use in the field.
	+ Note: This shall not hinder the progress in RAN4.
* Deprioritize Option 2 in view of Option 1 in RAN1 discussion for inter-vendor training collaboration
	+ Note: This deprioritization shall not affect the ongoing discussion in RAN4 on RAN4-Option3 and RAN4-Option4.
* Specify one or more Options among Option 3/4/5.

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Lenovo | We are still not clear on the feasibility of specifying a meaningful model for Option1.So, we suggest further discussion on that.* Further study the feasibility of ~~Specify, if feasible,~~ Option 1~~, at least for interoperability and RAN4 testing~~
	+ ~~RAN1 to further discuss whether specified model(s) from Option 1 is only for interoperability and RAN4 testing, in which case it can be handled by RAN4, or if the specified model(s) are also for use in the field.~~
	+ Note: This shall not hinder the progress in RAN4.
* Deprioritize Option 2 in view of Option 1 in RAN1 discussion for inter-vendor training collaboration
	+ Note: This deprioritization shall not affect the ongoing discussion in RAN4 on RAN4-Option3 and RAN4-Option4.
* Specify one or more Options among Option 3/4/5.

 |
| Huawei, HiSilicon | 1) For Option 1, the key issue is to clarify whether we only targets to ensure it support testing (Direction 1), or it supports usage in realistic network (Direction 2). For Direction 2, we may have further issues for further discussion to support its usage at realistic network2) For Option 2/3/4, it is too early to draw conclusion to specify them.* ~~Specify, if feasible, Option 1, at least for interoperability and RAN4 testing~~
* RAN1 to further discuss whether specified model(s) from Option 1 is
	+ Direction 1: only for interoperability and RAN4 testing, in which case it can be handled by RAN4, or
	+ Direction 2:~~if~~ the specified model(s) are also for use in the field.
		- Issue 2-1: whether the specified model is based on training data from field
		- Issue 2-2: whether the specified model targets to optimize the model structure for optimized performance.
		- Issue 2-3: whether to specify multiple models for model selection at different realistic scenarios
	+ Note: This shall not hinder the progress in RAN4.
* Deprioritize Option 2 in view of Option 1 in RAN1 discussion for inter-vendor training collaboration
	+ Note: This deprioritization shall not affect the ongoing discussion in RAN4 on RAN4-Option3 and RAN4-Option4.
* ~~Specify one or more Options among Option 3/4/5.~~
 |
|  |  |
|  |  |

Proposal 22a:

* For Option 1, it is clarified that NW/UE may either directly use the specified reference model(s) in implementation or use them only as reference to develop its own model(s) that match or are interoperable with the reference model(s).

|  |  |
| --- | --- |
| *Support / Can accept* | Lenovo (comment), Futurewei, New H3C |
| *Object / Have a concern* |  |

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Lenovo | We believe this is common to all options, not only option1, so we suggest:* ~~For Option 1,~~ it is clarified that NW/UE may either directly use the specified reference model(s) in implementation or use them only as reference to develop its own model(s) that match or are interoperable with the reference model(s).
 |
| Huawei, HiSilicon | OK. As a note, if they develop (and train) their own model(s), the testability may be questionable. |
| vivo | We would like to be more careful of the implication to implementation before we have a whole picture of options we choose.Reference model is a reference to implementation, that’s the current stage understanding. We do not need mix too many things at this stage. |

Proposal 23a:

* For Option 3/4/5, to simplify the discussion, focus further discussion on the model/parameter/dataset exchange originating from the NW-side and ending at the UE-side.
	+ Note: this is only for the purpose of simplified discussion and does not mean deprioritizing the exchange originating from the UE-side and ending at the NW-side from potential specification.
* For Option 3, further consider the two sub-options:
	+ 3a: Parameters received at the UE or UE-side goes through offline engineering at the UE-side for potential re-training, re-development of a different model, and/or offline testing.
		- The method of exchanging is either over the air-interface or offline.
		- Parameters exchanged from the NW-side to UE-side is either CSI generation or reconstruction part.
	+ 3b: Parameters received at the UE are directly used for inference at the UE in plug-and-play manner (assuming such UE capability).
		- The method of exchanging is over the air-interface (i.e., model transfer/deliver Case z1-z4).
		- Parameters exchanged from the NW-side to UE-side is CSI generation part.
* For Option 5, further consider the two sub-options:
	+ 5a: Model received at the UE or UE-side goes through offline engineering at the UE-side for potential re-training, re-development of a different model, and/or offline testing.
		- The method of exchanging is either over the air-interface or offline.
	+ 5b: Model received at the UE are directly used for inference at the UE in plug-and-play manner (assuming such UE capability).
		- The method of exchanging is over the air-interface (i.e., model transfer/deliver Case z1-z4).
		- Model exchanged from the NW-side to UE-side is CSI generation part.
* For Option 4:
	+ Dataset received at the UE or UE-side goes through offline engineering at the UE- side for re-training, re-development of a different model, and/or offline testing.
	+ The method of exchanging is either over the air-interface or offline.

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Lenovo | We generally support this proposal.We suggest adding the following point (under 3a) also to 5a and 4:* + - Parameters exchanged from the NW-side to UE-side is either CSI generation or reconstruction part.
 |
| Futurewei | In general, we are ok with this proposal. For Option 5b, we think the model exchange from NW-side to UE-side may also include the CSI reconstruction part for performance monitoring purpose as this is one of the options for UE-side performance monitoring as specified in TR 38.843 section 7.1.2. |
| Huawei, HiSilicon | OK with the direction. But some comments in below.1) For 3a/3b, the major difference is whether UE can run the delivered parameters in plug-and-play manner. For 3a, NW side cannot assume UE can run it in short time scale, while for 3b, NW can assume UE can run it in short time scale. Regarding how to process the parameters at the UE side (retrain/testing, etc.), it is UE side implementation.2) For Option 5, since Case z5 (over the air) has been deprioritized, we do not need to discuss it. So 5a over the air manner and 5b are removed.* For Option 3, further consider the two sub-options:
	+ 3a: Parameters received at the UE or UE-side without plug-and-play manner, i.e., long time scale before being able for inference
		- Note: Delivered parameter goes through offline engineering at the UE-side for potential re-training, re-development of a different model, and/or offline testing.
		- The method of exchanging is either over the air-interface or offline.
		- Parameters exchanged from the NW-side to UE-side is either CSI generation or reconstruction part.
	+ 3b: Parameters received at the UE are directly used for inference at the UE in plug-and-play manner (assuming such UE capability), i.e., short time scale before being able for inference.
		- The method of exchanging is over the air-interface (i.e., model transfer/deliver Case z1-z4).
		- Parameters exchanged from the NW-side to UE-side is CSI generation part.
* For Option 5, further consider the two sub-options:
	+ 5a: Model received at the UE or UE-side without plug-and-play manner, i.e., long time scale before being able for inference
		- Note: Delivered model goes through offline engineering at the UE-side for potential re-training, re-development of a different model, and/or offline testing.
		- The method of exchanging is ~~either over the air-interface or~~ offline.
	+ ~~5b: Model received at the UE are directly used for inference at the UE in plug-and-play manner (assuming such UE capability).~~
		- ~~The method of exchanging is over the air-interface (i.e., model transfer/deliver Case z1-z4).~~
		- ~~Model exchanged from the NW-side to UE-side is CSI generation part.~~
 |
| vivo | We are fine to further clarify different sub-options for each option.Several general comments:* There are ambiguities related to the terminology “plug and play”. We would like to use a more generic term, e.g., “on device operation”. Or “short time scale” is also fine.
* We would like to make the division of sub-options simplified: with offline engineering or with “on device operatoin”, without involving too many details.
* For Option3 over air interface, our understanding is it is model transfer case z4. For Optoin5 over air interface, our understanding is it is model transfer case z3. We don’t think there is case z1 or z2 involved.

Based on such spirit, we would like to update the original categorization into the following:* For Option 3, further consider the two sub-options:
	+ 3a: Parameters received at the UE or UE-side goes through offline engineering at the UE-side assuming long time scale ~~for potential re-training, re-development of a different model, and/or offline testing.~~
		- The method of exchanging is either over the air-interface or offline.
		- Parameters exchanged from the NW-side to UE-side is either CSI generation or reconstruction part.
	+ 3b: Parameters received at the UE are directly used for inference at the UE assuming on device operation/short time scale ~~in plug-and-play manner~~ (assuming such UE capability).
		- The method of exchanging is over the air-interface (i.e., model transfer/deliver Case ~~z1-~~z4).
		- Parameters exchanged from the NW-side to UE-side is CSI generation part.
* For Option 5, further consider the two sub-options:
	+ 5a: Model received at the UE or UE-side goes through offline engineering at the UE-side assuming long time scale (e.g. days) ~~for potential re-training, re-development of a different model, and/or offline testing.~~
		- The method of exchanging is either over the air-interface or offline.
	+ 5b: Model received at the UE are directly used for inference at the UE assuming on device operation/short time scale ~~in plug-and-play manner~~ (assuming such UE capability).
		- The method of exchanging is over the air-interface (i.e., model transfer/deliver Case z3~~1-z4~~).
		- Model exchanged from the NW-side to UE-side is CSI generation part.
* For Option 4:
	+ Dataset received at the UE or UE-side goes through offline engineering at the UE- side ~~for re-training, re-development of a different model, and/or offline testing.~~
		- The method of exchanging is either over the air-interface or offline.
 |

Proposal 24a:

For Option 3a/4/5a:

* Study different methods of exchanging the parameters/dataset/model that satisfy all the following, including over the air-interface and offline delivery:
	+ Acceptable inter-vendor collaboration complexity
	+ Scalabile at the NW side (e.g., UE-vendor-agnostic CSI reconstruction model)
	+ Feasible for UE implementation, potentially with intra-vendor offline engineering
* Study methods to ensure/improve interoperability and good field performance.
	+ e.g., consider exchanging the target performance, along with the model/dataset/parameters exchange.

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Lenovo | Support |
| Futurewei | We are ok with this proposal in general, we suggest some wording change:“Study different methods of exchanging the parameters/dataset/model that best satisfy ~~all~~ the following, including over the air-interface and offline delivery”.There is a typo in the second bullet: “Scalable”. |
| Huawei, HiSilicon | Can FL clarify the intention/meaning of these two bullets?* + Scalabile at the NW side (e.g., UE-vendor-agnostic CSI reconstruction model)
	+ Feasible for UE implementation, potentially with intra-vendor offline engineering
 |
| vivo | Seems all these bullets are within what we need to study based on previous agreement?**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.*** |

Proposal 25a:

For Option 1/2/3, study how to specify the model structure (Option 1/3), dataset (Option 2), and/or parameters (Option 1). Study more details at least on the following aspects:

* Whether to specify CSI generation part or CSI reconstruction part or both
* Evaluation assumptions and/or field dataset
* Input/output data format and loss function
* Model architecture candidates
* Considerations for generalization and scalability
* Procedure/methodology for RAN1 to agree on dataset, model, and training for specification.

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

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Lenovo | We suggest first to down-select options and then if needed go to details of each options. |
| Huawei, HiSilicon | Fine with the direction for discussion.We need to point out that the discussion for Option 1/2 may additionally require the alignment of dataset and training related details, while Option 3 may not need to align them if companies can cross check other architectures internally.For Option 1/2/3, study how to specify the model structure (Option 1/3), dataset (Option 2), and/or parameters (Option 1). Study more details at least on the following aspects:* Whether to specify CSI generation part or CSI reconstruction part or both
* Evaluation assumptions and/or field dataset
* Input/output data format/type ~~and loss function~~
* Quantization method, e.g., scalar quantization, vector quantization.
* Model architecture candidates
* Layer related parameters
* Considerations for ~~generalization and~~ scalability
* Procedure/methodology for RAN1 to agree on dataset, model, and training for specification.
* In particular, at least for Option 1/2, following aspects are studied in addition:
	+ Considerations for generalization
	+ Common dataset
	+ Quantization awareness for training, e.g., training aware Case 2-1 or Case 2-2
	+ Training parameters, e.g., batch size, epoch number, learning rate, loss function, training completion criteria, etc.
 |
| vivo | Seems Option2 does not need to be further studied based on previous FL proposal on deprioritizing Option2. |

### Others

Please provide any other comments regarding inter-vendor collaboration.

|  |  |
| --- | --- |
| *Company* | *Comments* |
| Huawei, HiSilicon | As Option 1/3 need standardized model/model structure, we need a large work load of simulation and cross check among companies. To facilitate companies to study the way for converging on the model/model structure, we may need to agree on the principles (as the elaboration of “Procedure/methodology for RAN1 to agree on dataset, model, and training for specification” in Proposal 25a). E.g.,How to generate the dataset – 1) common, or up to companies. 2) SLS, LLS, or field data. 3) whether to calibrate the channel model for data generation.How to converge on the model – companies to directly bring full model structure for cross check, or the RAN1 group first agrees on the backbone then layer by layer.What are the metrics to down select the model structure, possibly as a tradeoff between complexity and performance.For Option 4, similarly, there are also some issues for further study, e.g., Data format: 1) The type of ground-truth CSI (eigenvectors, channel matrix, angular-delay domain eigenvectors, etc.), the dimension of input (port number, bandwidth, etc.)/output (CSI payload size). 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.For the above issues, we may need to trigger the discussions on the potential issues and solutions, and then do necessary down selection before Sept. |
|  |  |

# Data collection

## Summary of company proposals

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

Huawei

Proposal 8: 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 9: 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 10: 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.

Intel

Proposal 3: Whether/how to support new CSI report format for ground truth CSI quantization should be further studied considering the corresponding CSI overhead.

ZTE

Proposal 10: 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 11: 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

CATT

Proposal 9: In CSI compression using two-sided model use case, discuss data collection for training as if it would be specified in Rel-19.

Proposal 10: 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 3: 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 7: 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 8: 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 9: 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.

Xiaomi

Proposal 1: 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.

Fujitsu

Proposal 12: 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.

Google

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

Lenovo

Proposal 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.

Proposal 2: Support procedures/signaling enabling UE/NW for transmission of subset of samples among the set of measured/collected samples from the environment.

Proposal 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.

## Discussion

|  |  |
| --- | --- |
| Companies | Views |
| Huawei | High resolution eT2, L=8, 10, 12, pv = 0.8, 0.9, 0.95, ref amplitude=4, phase = 5 / 6Rank configured by NWFor Training Type3, study dataset ID used to differentiate model trained at opposite side; dataset sizeNW side send dataset in different portions for different UE |
| Intel | New CSI report format for ground-truth quantization |
| ZTE | High-resolution eT2UE report SINR, CQI, or with configured data quality when upload data to NW |
| CATT | Procedure: NW collection, UE collectionContent: data format, typeAssistance informationEnhancement of CSI-RS configuration for data collection |
| China Telecom | SINR, CQI or with configured data quality when upload data to NW |
| CMCC | High resolution eT2 or R18 Doppler eT2 |
| Xiaomi | Study overhead reduction, no latency requirement |
| Fujistu | High resolution eT2 |
| Google | Data-collection mechanism: NW configuration or UE request |
| Lenovo | High resolution eT2Association between additional conditions with data samples |

Companies brought discussions on

* ***Data collection procedures and mechanisms***
* ***High-resolution codebook-based report of ground-truth***
* ***Associated assistance information and dataset ID***
* ***Indication or configuration of dataset quality***
* ***CSI-RS enhancement***

However, this discussion can wait a bit, and in this meeting we can prioritize discussion on aspects that are more crucial toward September checkpoint.

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

# Monitoring

## Summary of company proposals

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

Huawei

Proposal 11: 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 12: There is no strong motivation for specifying the UE side proxy model for monitoring.

Proposal 13: 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

Proposal 4: 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 5: 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.

ZTE

Proposal 13: 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 14: In CSI compression using two-sided model use case, deprioritize the study on UE-side monitoring in Rel-19 study phase.

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

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

CATT

Proposal 11: 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

Proposal 5: 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 6: 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 6: For performance monitoring, the following two options could be prioritized:

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

Xiaomi

Proposal 2: 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 4: The privacy/proprietary should be considered for the comparison between dataset delivery (Case 4) and model delivery (Case 3/5).

Proposal 5: Support NW-side monitoring based on the target CSI with realistic channel estimation associated to the CSI report.

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

Proposal 7: Support UE-side monitoring based on the output of the CSI reconstruction model.

Proposal 8: 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.

Fujitsu

Proposal 19: 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 20: 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 21: 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 22: 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 23: 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 24: 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.

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

Apple

Proposal 1: 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 2: For CSI compression using two-sided model, for UE side performance, further study RLF like mechanism for UE initiated report.

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 and eventual KPI can be considered
* Reporting mode (per sample or statistic)
* Deprioritize input/output-based and legacy CSI based monitoring
* Signaling of ground-truth CSI and reconstructed CSI between two sides
 |
| Intel | * Using SRS as ground-truth
 |
| ZTE | * UE report ground-truth generated from realistic channel estimation
* Deprioritize UE side monitoring
 |
| Google | * Not use SGCS for monitoring
* use hypo BLER instead, BLER based on the CQI from non-ML codebook and the ML CSI.
* No mandate of UE supporting eT2
 |
| CATT | * Deprioritize proxy-based monitoring
 |
| China Telecom | * Prioritize ground-truth reporting
* Deprioritize reconstructed CSI indication from NW to UE
 |
| CMCC | * Prioritize ground-truth reporting and reconstructed CSI from NW to UE
 |
| Xiaomi | * Using intermediate KPI or using existing CSI as reference
 |
| NEC | * Ground-truth reporting
* Reconstructed CSI indication from NW to UE or UE having the CSI reconstruction model
 |
| Fujistu | * Proxy model: feasibility, generalization ability, reliability, should be studied before move on
* Prioritize ground-truth reporting
* For NW side monitoring using legacy CSI as reference, study 1) UE selects and report PMI; 2) UE computes and reports intermediate KPI for the ref scheme; 3)NW selects the PMI based on ground-truth CSI reported from UE
* Study follow-up actions
* Study monitoring of an inactive model (its initial activation, and re-activation)
 |
| InterDigital | * Mechanisms, time-based or event-triggered
* Proper UE-side monitoring metric and its reporting
* NW-side with lower signaling overhead
 |
| Apple | * NW implicitly transmits reconstructed CSI via precoded CSI-RS w/ hypo BLER as metric
* RLF like mechanism for UE initiated report
 |

Companies brought discussion on monitoring

* Performance monitoring mechanisms (e.g., time-based or event-triggered / UE initiated, etc)
* Performance monitoring reference and metrics (e.g., SGCS, or hypothetical BLER)
* NW-side monitoring based on
	+ ground-truth reporting from UE side using existing codebook (e.g., Type I, eType II)
	+ SRS measurement
* UE-side monitoring and reporting of monitoring results / metrics with following options
	+ Option 1: NW indication of reconstructed CSI
	+ Option 2: via precoded CSI-RS using reconstructed CSI
	+ Option 3: UE-side running CSI reconstruction model or its proxy model
	+ Option 4: UE-side running a direct intermediate KPI estimator (e.g., SGCS estimator)

This discussion can wait a bit, while we first focus on inter-vendor training collaboration aspects and evaluation for temporal aspects and localized models.

|  |  |
| --- | --- |
| *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 14: 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 15: 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 16: 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 17: 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.

ZTE

Proposal 12: For CQI determination, at least prioritize the specification impact discussions on Option 1a, Option 1b.

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 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 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

CATT

Proposal 12: In CSI compression using two-sided model use case, standardize quantization scheme.

Proposal 13: In CSI compression using two-sided model use case, legacy CSI reporting principles is reused as much as possible.

Proposal 14: 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 15: 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 4: For CSI compression sub use case, the pairing information should be included in the process of functionality/model identification.

Xiaomi

Proposal 3: The legacy priority rule can be reused to define the priority the AI/ML based CSI reporting, and a priority value $Pri\_{iCSI}\left(y,k,c,s\right)$$Pri\_{iCSI}\left(y,k,c,s\right)$ $Pri\_{iCSI}\left(y,k,c,s\right)$with new parameter value or introducing new parameter $Pri\_{iCSI}\left(y^{'},y,k,c,s\right)$ is used to indicate the priority of AI/ML based CSI reporting.

Proposal 4: 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 5: 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 6: 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 9: 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 10: 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.

Fujitsu

Proposal 13: 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 14: 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 15: 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 16: 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 17: 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 18: 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.

Lenovo

Proposal 13: Support definition of pairing information based on the conditions/additional conditions assigned to the samples of the datasets used for training of the model.

Proposal 14: Further study model identification/selection procedures during inference time when different models have been developed for different UE-NW vendor pairs.

Proposal 15: Support procedures/signalling enabling CSI-compression models having both Scaler and vector Quantizers for generation of the CSI-feedback bits

Nokia

Proposal 10: 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 11: RAN1 to study the feedback of CQI for different rank hypotheses.

Proposal 12: 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.

CEWiT

Proposal-8: Consider using model ID based identification for ensuring proper training between UE sided model and NW sided model

Proposal-9: 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-10: 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).

## Discussion

|  |  |
| --- | --- |
| Companies | Views |
| Huawei | * Quantization: segmentation for VQ, configuration/updating/reporting mechanism, configuration of granularity/range for SQ
* Downselection of pairing options
* CQI: option 1 (not based on reconstructed CSI)
* CPU, priority rule, CSI mapping (two-part UCI)
 |
| ZTE | * Prioritize option 1a (target CSI) / 1b (target CSI w/ adjustment)
 |
| Google | * Support precoder feedback, channel feedback and W2 feedback
* Priority rule: non-AI > AI
* CPU occupancy: type1 (MPU) for channel estimation and pre-processing; type2 (IPU) for ML inference
* Reusing existing CPU framework
* Hybrid CSI: high rank for Type I, low rank for ML
 |
| CATT | * Quantization specified
* CQI option 1a / 1b / 1c
 |
| China Telecom | * Pairing information included in functionality / model identification
 |
| Xiaomi | * New parameter in priority rule for ML CSF
* UCI part divided into N groups for omission
 |
| NEC | * CQI option 2a if reconstruction model is at UE, otherwise support option 1a/1b
* Pairing: consider downselection from option 1 (decoder ID), option 2 (encoder IE), paring ID
 |
| Fujistu | * CSI format: SF domain or angular-delay domain
* Pairing procedure: NW initiated or UE initiated
* Global model is sufficient, no need of pairing ID, local ID can be considered to reduce signalling overhead
 |
| Lenovo | * Defining pairing information based on conditions / additional conditions assigned to datasets
* Study model identification / selection procedure during inference when different models have been developed for different UEs
* Signalings / procedures enabling CSI encoder having both VQ and SQ
 |
| Nokia | * CQI: focus on option 1a / 2a-1, also consider 1b / 2a-2
* Study CQI for different rank
* Study spec efforts for rank-specific/common, layer-specific/common architectures
 |
| CEWiT | * Model id for identification
* Model pairing before inference w/ assistance of UE capability report to ensure NW side model can avoid any model mismatch
* 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)
 |

### Pairing

Several companies proposed to continue study of model identification and model pairing procedure with more details and/or down-selection (6): Huawei, China Telecom, NEC, Fujistu, Lenovo, CEWiT

This is related to inter-vendor training collaboration Options, so we can wait for the progress of inter-vendor collaboration Options.

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

### CQI determination in CSI report

Several companies brought discussion on CQI determination with preferences:

CQI options (5):

* prioritize option 1 family (Huawei, ZTE, CATT, NEC, Nokia);
* Also okay for option 2a (NEC, Nokia)

This discussion can wait a bit, while we first focus on inter-vendor training collaboration aspects and evaluation for temporal aspects and localized models.

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

### Options for rank>1 solution

This discussion can wait a bit, while we first focus on inter-vendor training collaboration aspects and evaluation for temporal aspects and localized models.

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

### Quantization

Specification of quantization (2): Huawei, CATT

This discussion can wait a bit, while we first focus on inter-vendor training collaboration aspects and evaluation for temporal aspects and localized models.

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

### Other inference aspects

CPU occupancy, priority rule, UCI mapping (3): Huawei, Google, Xiaomi

We can discuss other higher priority topics first.

|  |  |
| --- | --- |
| *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#10: Consider to study joint source-coding, channel-coding and modulation (JSCCM)-based CSI feedback for CSI compression.

Ericsson

[Proposal 3: Consider assuming antenna arrays where the NR codebook design assumption of uniform planar 2D antenna array of equally spaced identical subarrays doesn’t hold for the CSI compression use case.](#_Toc163231812)

Beijing Jiaotong University

Proposal 1: Consider to study joint source-channel coding (JSCC) based framework for CSI feedback..

Proposal 2: Study symbol-level DL-based CSI feedback as a use case for CSI feedback enhancement..

Proposal 3: The channel recovery capability of the symbol-level DL-based CSI feedback better than the bit-level DL-based CSI feedback.

## Discussion

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

# Proposals for offline/online sessions

## Proposals for Monday 4/22 GTW session

# FL closing remark

TBD

# 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

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