**3GPP TSG RAN WG2 Meeting #129bis R2-250xxxx**

Wuhan, China, Apr 7th– 11th, 2025

Agenda Item: 8.1.x

Source: Xiaomi, Ericsson

Title: Report of [POST129][029][AI Phy] Model transfer (Xiaomi/Ericsson)

Document for: Discussion and Decision

# Introduction

This report provides a summary for the following post-meeting email discussion:

* [POST129][029][AI Phy] Model transfer (Xiaomi/Ericsson)

Intended outcome: Identify the options for OTA and non-OTA, based on TR, contributions and considering data collection discussion.

Deadline: long

Considering this is the first time in Rel-19 we discuss different solutions in OTA and non-OTA, rapporteurs suggest to have two phases:

**Phase 1:** Solution identification, illustration and Q&A among companies to reach consensus. Focusing on technical discussion on how each solution works.

Deadline for providing comments for phase 1 is March 11th, 2025, 10:00UTC.

**Phase 2:** Based on solutions identified during Phase 1, companies are welcomed to provide further comment on complexity and feasibility analysis.

Deadline for providing comments for phase 2 is March 21st, 2025, 10:00UTC.

Companies providing input to this email discussion are invited to leave contact information below.

|  |  |  |
| --- | --- | --- |
| **Company** | **Name** | **Email Address** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# Phase 1 Discussion

## Evaluation Area/Requirement

During Rel-18 SI, we had some practices on how to evaluate different model transfer/delivery solutions among following four discussion areas:

- A1: Large, no upper limit model/model parameter size,

- A2: Model transfer/delivery continuity (i.e., resume transmission of model (segments) across gNBs),

- A3: Network controllability on model transfer/delivery (e.g., management decision at gNB),

- A4: Model transfer/delivery QoS (for DRB) (including latency, etc.) and priority (for SRB).

In RAN2 #129 meeting, following requirements are proposed/summarized from T-mobile, etc [0949]:

1. Low priority/QoS than user traffic (A4)
2. NW controllability: 1) if and when to transfer/delivery the model securely in a NW-aware manner (A3)
3. Model visibility: addressable model that UE can request for a specific model
4. Initiation: initiated by a UE

Additionally, CMCC, etc [1051] further discussed the visibility and controllability of two-sided model:

1. Model visibility: open format and known structure, where parameters are transferred from NW to UE
2. NW controllability: whole model is trained at NW (A3)

Furthermore, according to RAN1 LS R2-2500015, following model parameter and/or dataset size can be summarized as below:

Option 4-1 (sharing {target CSI, CSI feedback} dataset): around 225MB

Option 3a-1 without target CSI (sharing encoder parameter): ranging from 36KB to 52MB, 11.6MB in average

Option 3a-1 with target CSI (sharing encoder parameters, along with {target CSI} dataset): 225MB + 11.6MB in average

In the end, RAN1 also mentioned below understanding of latency and frequency of dataset and/or parameter sharing:

|  |
| --- |
| Regarding latency and frequency of the dataset and/or parameter sharing: As the purpose of the dataset and/or parameter sharing is for a UE-side offline training, the dataset and/or parameter sharing is expected to have relaxed latency (e.g., days/weeks) and be infrequent. |

Based on above information, rapporteurs try to summarize the following discussion area/requirements for evaluation of model transfer/delivery solutions:

* A1: Minimum dataset and/or parameter sharing size can be 36kB. In average, dataset and/or parameter sharing size can be as large as 225MB+11.6MB;
* A2: Model transfer/delivery continuity needs to be supported considering dataset and/or parameter sharing may be expected to transfer in days/weeks;
* A3: NW controllability: Decision on if and when to transfer/delivery the dataset and/or model parameter securely in a NW-aware manner;
* A4: Low priority/QoS than user traffic, with relaxed latency requirement and infrequent update;
* A5: Model visibility: open format and known structure.

##### Q1-1. Do you agree the above discussion areas/requirements for two-sided model transfer/delivery solution evaluation? (Please see Q1-2 for new discussion areas/requirements)

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comment (if No, please comment with expected requirement for the corresponding discussion area)** |
| [Example] | A1: Yes  A2: No  A3: Yes  A4: Yes  A5: No | A2: [Comment and expected requirement for the corresponding discussion area]  A5: [Comment and expected requirement for the corresponding discussion area] |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

##### Q1-2: Any other discussion areas/requirements for two-sided model transfer/delivery solution evaluation?

|  |  |
| --- | --- |
| **Company** | **Comment** |
|  |  |
|  |  |
|  |  |
|  |  |

##### Q1-3: Any questions would like to ask RAN1 for further clarification?

|  |  |
| --- | --- |
| **Company** | **Comment** |
|  |  |
|  |  |
|  |  |
|  |  |

## Candidate Solutions

According to contributions submitted to RAN2 #129 meeting, it seems companies have different understanding on the termination of model transfer/delivery (e.g. UE or UE-side OTT server). Before discussing candidate solutions for model transfer/delivery, rapporteurs think it would be good to first clarify the discussion scope and background based on RAN1 LS.

Since RAN1 #116 meeting, RAN1 has been discussing model transfer/delivery methods for CSI compression, where Option 1-5 were identified and analysed.

|  |
| --- |
| * ***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*** |

According to RAN1 discussion till RAN1 #118bis meeting, following options can be summarized, where the solutions that are still on the table are highlighted in green:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Options** | **Direction** | **Information for model transfer/delivery** | | **Requirement of offline training?** |
| Option 1 | C | Fully standardized reference model (structure + parameters) | |  |
| Option 2 | C | Standardized dataset | |  |
| Option 3a | A | Standardized reference model structure + **Parameter exchange** | 3a-1: CSI generation part (with/without target CSI) | Offline engineering at UE-side OTT server |
| 3a-2: CSI reconstruction part |
| 3a-3: both parts |
| Option 3b (z4) | B |  | On-device operation without offline engineering |
| Option 4 | A | Standardized data / dataset format + **Dataset exchange** | 4-1: target CSI, CSI feedback | Offline engineering at UE-side OTT server |
| 4-2: CSI feedback, reconstructed target CSI |
| 4-3: target CSI, CSI feedback, reconstructed target CSI |
| Option 5a | A | Standardized model format + Reference model exchange | | Offline engineering at UE-side OTT server |
| Option 5b (z4) | B | On-device operation without offline engineering (model structure is aligned based on offline inter-vendor collaboration) |

It is clear from RAN1 LS that option 3a-1 (with/without target CSI) and option 4-1 are within Direction A, where offline engineering at UE-side OTT server is required.

Observation #: Option 3a-1 (with/without target CSI) and Option 4-1 are within Direction A, where offline engineering at UE-side OTT server is required.

According to contributions submitted to RAN2 #129 meeting (e.g. MTK[0323], Ericsson[1288], HW[1111], Apple[0263], vivo[0128], QC[0394], Lenovo[0614], ZTE[0836], SS[0910], Nokia[0998]), following two alternatives can be further considered as model transfer path:

**Alternative 1 (non-OTA approach):**

**gNB** -> **NW dataset/model parameters collection entity** -> **UE training entity** (OTT server inside/outside of MNO) -> **UE** (UE model parameter delivery for inference)

|  |
| --- |
| base station, cell tower, communication, connection, network, tower cloud, database, hosting, server iphone 14, iphone, mobile, smartphone, device, app, pro cloud, server, web  model transfer for UE inference  dataset/model parameters transfer  dataset/model parameter transfer  CSI compression data collection at gNB  NW-side dataset/model parameters collection entity (gNB/CN/OAM/gNB server) for two-sided model training  UE-side training entity for two-sided UE part model training |

**Alternative 2 (OTA approach):**

**gNB** -> **NW dataset/model parameters collection entity** -> **UE** -> **UE training entity** (OTT server inside/outside of MNO) -> **UE** (UE model parameter delivery for inference)

|  |
| --- |
| cloud, server, web iphone 14, iphone, mobile, smartphone, device, app, pro cloud, database, hosting, server base station, cell tower, communication, connection, network, tower  dataset/model parameter for training  dataset/model parameters transfer  dataset/model parameter transfer  NW-side dataset/model parameters collection entity (gNB/CN/OAM/gNB server) for two-sided model training  UE-side training entity for two-sided UE part model training  CSI compression data collection at gNB  dataset transfer  model transfer for UE inference |

The identified transfer path will be further discussed in details (e.g. either be standardized or by implementation (e.g. outside of 3GPP)) in the following questions.

##### Q2-0: Do companies agree with the above two alternatives of model transfer/delivery (no matter by implementation or standardization)? Note that the intermediate nodes between each entity (e.g. NW dataset/model parameters collection entity <-> UE training entity, UE <-> UE training entity) will be further discussed in Section 2.2.1 and Section 2.2.2.

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comment (intermediate nodes will be discussed in later questions, this question only focuses on the overall path)** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

One common transmission path of two alternatives is dataset transfer from gNB to NW dataset/model parameters collection entity (e.g. gNB/OAM/CN/gNB server).

Furthermore, it was captured in TR38.843:

|  |
| --- |
| - Model Training:  o For the two-sided CSI compression use case, training data can be generated by either the UE or the gNB, depending on specific requirements, while the termination point for training data may include the gNB, OAM, Over-The-Top (OTT) server or UE.  § Note: RAN2 identified the case in which Core Network may be used for model training. However, no study was conducted since this is beyond the scope of this Working Group. |

If OAM/CN are involved as NW dataset/model parameters collection entity, this transfer path seems to fall into SA2/SA5 scope, which is outside of RAN2 scope.

##### Q2-1: Do companies agree that transfer path from gNB to NW dataset/model parameters collection entity (OAM/CN/gNB server), if needed, is up to SA2/SA5?

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comment** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

### Non-OTA approach

In this section, let’s focus on how to transfer dataset/model parameters between each identified entity.

#### NW dataset/model parameters collection entity -> UE side OTT server (inside/outside MNO)

After receiving training data collection from gNB, NW dataset/model parameters collection entity may further transfer dataset/model parameters used for two-side model UE-part training to UE-side OTT server.

It was proposed in MTK[0323], Ericsson[1288], Xiaomi[0265], QC[0394], ZTE[0836], SS[0910], Nokia[0998] that model parameter/dataset can be transferred from NW dataset/model parameters collection entity to UE-side OTT server, where UE-side OTT server may be either inside or outside of MNO.

Following options were proposed from contributions, rapporteurs further provide impacted WGs and specification/implementation impact as below:

|  |  |  |
| --- | --- | --- |
| **Option** | **Impacted WG** | **Specification impact/Implementation impact** |
| 1. gNB -> OAM -> UE-side OTT server | SA5 | Up to SA5 |
| 1. gNB -> CN -> UE-side OTT server | RAN3, SA2 | Up to RAN3 on NG impact, SA2 |
| 1. gNB -> UE-side OTT server (outside of MNO) | Outside of 3GPP | Up to implementation |
| 1. gNB -> server inside MNO -> optionally OTT server (outside of MNO) | SA2 | Up to SA2 |

##### Q2-2: Do companies agree with above analysis on specification/implementation impact and impacted WGs? Proponent companies are also welcomed to add specification/implementation impact.

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comment** |
| [Example] | 1. Yes 2. Yes 3. No 4. Yes | Option 3) ‘specification impact’: [comment/new impact] |
|  |  |  |
|  |  |  |
|  |  |  |

##### Q2-3: Any missing options? If yes, please explain the proposed transfer path, and specification impact/implementation impact/impacted WG.

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Transfer path** | **Specification impact/Implementation impact** | **Impacted WG** |
|  |  |  |  |
|  |  |  |  |

#### UE side OTT server -> UE

After receiving training dataset or model parameter, UE side OTT server starts to retrain two-sided model UE-part, and UE side OTT server further transfers/deploys model (parameter) to UE directly for model inference.

MTK[0323] proposes UE-side OTT server transfers model (parameter) to UE via OAM and gNB. On the other hand, it is also possible that the transfer path from OTT server to UE can be left to implementation.

Note that RAN2 studied 8 solutions to support model transfer/delivery from network to UE via CP/UP:

- Solution 1a: gNB can transfer/deliver AI/ML model(s) to UE via RRC signalling.

- Solution 2a: Core Network (except LMF) can transfer/deliver AI/ML model(s) to UE via NAS signalling.

- Solution 3a: LMF can transfer/deliver AI/ML model(s) to UE via LPP signalling.

- Solution 1b: gNB can transfer/deliver AI/ML model(s) to UE via UP data.

- Solution 2b: Core Network (except LMF) can transfer/deliver AI/ML model(s) to UE via User Plane (UP) data.

- Solution 3b: LMF can transfer/deliver AI/ML model(s) to UE via UP data.

- Solution 4a: OTT server can transfer/deliver AI/ML model(s) to UE (e.g., transparent to 3GPP).

- Solution 4b: OAM can transfer/deliver AI/ML model(s) to UE.

Rapporteurs further provide potential options and impacted WGs, specification/implementation impact as below:

|  |  |  |
| --- | --- | --- |
| **Option** | **Impacted WG** | **Specification impact/Implementation impact** |
| 1. UE-side OTT server -> UE | Outside of 3GPP | Solution 4a  [the detailed specification impact of Solution 4a is captured in TR38.843] |
| 1. UE-side OTT server -> OAM -> gNB -> UE | SA5 | UE-side OTT Server -> OAM is up to SA5,  OAM -> UE follows Solution 4b  [the detailed specification impact of Solution 4b is captured in TR38.843] |
| 1. UE-side OTT server -> CN -> gNB -> UE | SA2 | UE-side OTT Server -> CN is up to SA2,  CN -> UE follows Solution 2a/2b  [the detailed specification impact of Solution 2a/2b is captured in TR38.843] |

##### Q2-4: Do companies agree with above analysis on specification/implementation impact and impacted WGs? Proponent companies are also welcomed to add specification/implementation impact.

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comment** |
| [Example] | 1. Yes 2. Yes 3. No | Option 3) ‘specification impact’: [comment/new impact] |
|  |  |  |
|  |  |  |
|  |  |  |

##### Q2-5: Any missing options? If yes, please explain the proposed transfer path, and specification impact/implementation impact/ impacted WG.

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Transfer path** | **Specification impact/Implementation impact** | **Impacted WG** |
|  |  |  |  |
|  |  |  |  |

### OTA approach

#### gNB -> UE (direct)

In this transfer path, gNB directly transfers the dataset and/or model parameter to UE (the dataset/model parameter will be further propagated to UE-side OTT server in Direction A). During Rel-18 SI, RAN2 identified Solution 1a and Solution 1b (and corresponding specification impact) as candidate solutions:

1) Solution 1a: gNB can transfer/deliver AI/ML model(s) to UE via RRC signalling.

2) Solution 1b: gNB can transfer/deliver AI/ML model(s) to UE via UP data.

Note that, according to RAN1 LS, minimum dataset and/or parameter sharing size can be 36kB. In average, dataset and/or parameter sharing size can be as large as 225MB+11.6MB.

Furthermore, it is also rapporteurs’ understanding that this approach may also be appliable for Option 3b ‘Standardized reference model structure + Parameter exchange’ (On-device operation without offline engineering).

##### Q2-6: Do companies agree that Solution 1a and Solution 1b to be considered as candidate solution of ‘gNB -> UE’ of OTA approach? The specification impact of Solution 1a and Solution 1b in TR38.843 can be reused.

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comment** |
|  |  |  |
|  |  |  |
|  |  |  |

##### Q2-7: Any missing options? If yes, please explain the proposed transfer path, and specification impact/implementation impact/ impacted WG.

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Transfer path** | **Specification impact/Implementation impact** | **Impacted WG** |
|  |  |  |  |
|  |  |  |  |

##### Q2-8: Do companies agree that the identified solution for gNB -> UE (direct) is also applicable for Option 3b ‘Standardized reference model structure + Parameter exchange’ (On-device operation without offline engineering)?

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comment** |
|  |  |  |
|  |  |  |
|  |  |  |

#### NW dataset/model parameters collection entity -> UE

Instead of transmitting dataset/model parameter directly from gNB to UE, depending on which node (gNB/CN/OAM) is in charge of collecting the NW dataset/model parameters, the transfer path could be different:

|  |  |  |
| --- | --- | --- |
| **Option** | **Impacted WG** | **Specification impact/Implementation impact** |
| 1. gNB -> UE | Same as gNB -> UE direct link | |
| 1. CN -> gNB -> UE | RAN3, SA2 | Solution 2a/2b  [the detailed specification impact of Solution 2a/2b is captured in TR38.843] |
| 1. OAM -> gNB -> UE | SA5 | Solution 4b  [the detailed specification impact of Solution 4b is captured in TR38.843] |

##### Q2-9: Do companies agree with above analysis on specification/implementation impact and impacted WGs? Proponent companies are also welcomed to add specification/implementation impact.

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comment** |
| [Example] | 1. Yes 2. Yes 3. No | Option 3) ‘specification impact’: [comment/new impact] |
|  |  |  |
|  |  |  |
|  |  |  |

##### Q2-10: Any missing options? If yes, please explain the proposed transfer path, and specification impact/implementation impact/impacted WG.

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Transfer path** | **Specification impact/Implementation impact** | **Impacted WG** |
|  |  |  |  |
|  |  |  |  |

#### UE -> OTT server (similar to UE-side data collection)

As discussed at the beginning of Section 2.2, in Direction A, offline training for UE-part two-sided model takes place at UE-side OTT server. Therefore, in OTA approach, UE needs to further propagate the received dataset/model parameters to UE-side OTT server.

During previous RAN2 discussion, RAN2 has concluded following solutions for UE-side data collection:

1. UE collects and directly transfers training data to the Over-The-Top (OTT) server;

1a) OTT (Transparent)

1b) OTT (non-Transparent)

2. UE collects training data and transfers it to Core Network. Core Network transfers the training data to the OTT server via CP/UP.

3. UE collects training data and transfers it to OAM. OAM transfers the needed data to the OTT server.

Rapporteurs believe that this transfer path can reuse the same solution as UE-side data collection.

##### Q2-11: Do companies agree the above solutions for UE-side data collection and the corresponding analysis can also be used for the model transfer path UE -> OTT server?

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comment** |
|  |  |  |
|  |  |  |
|  |  |  |

#### UE-side OTT server -> UE

This transfer path is similar to OTT server -> UE in non-OTA approach, which is discussed in Section 2.2.1. We will not repeat the discussion here.

# Phase 2 Discussion

After phase 1 discussion, rapporteurs believe companies now have better understanding on how each solution works. During phase 2 discussion, let’s focus on the complexity and feasibility analysis of each solution.

# Conclusion

# Reference

[1] R2-2500323 Feasibility Analysis on RAN1 Identified Solution for Two-sided Model MediaTek Inc. discussion

[2] R2-2501288 On signalling feasibility of dataset and parameter sharing Ericsson discussion

[3] R2-2501111 Discussion on signalling feasibility of dataset and parameter sharing for CSI compression Huawei, HiSilicon discussion Rel-19 NR\_AIML\_air-Core

[4] R2-2500263 Discussion on parameters/model transfer in two-sided model Apple discussion Rel-19 NR\_AIML\_air-Core

[5] R2-2500296 Signalling feasibility of dataset and parameter sharing NEC discussion Rel-19 NR\_AIML\_air-Core

[6] R2-2500949 Requirements for Model Transfer/Delivery T-Mobile USA Inc., Boost Mobile Network, Deutsche Telekom, Orange, Charter Communication, Nokia Corporation discussion Rel-19 NR\_AIML\_air-Core

[7] R2-2501051 Discussion on AIML model transfer delivery CMCC,China Unicom,China Telecom,CATT,ZTE,Apple,Samsung discussion Rel-19 NR\_AIML\_air-Core

[8] R2-2501215 Discussion on model transfer/delivery NTT DOCOMO, INC. discussion Rel-19

[9] R2-2500128 Discussion on signaling feasibility of dataset and parameter vivo discussion NR\_AIML\_air-Core

[10] R2-2500156 Open Discussion on Two Sided Model OPPO discussion Rel-19 NR\_AIML\_air-Core

[11] R2-2500242 Signalling feasibility of AIML model transfer CATT discussion Rel-19 NR\_AIML\_air-Core

[12] R2-2500265 Feasibility analysis of model/dataset transfer solutions Xiaomi discussion Rel-19 NR\_AIML\_air-Core

[13] R2-2500394 Discussion on Dataset and Parameter Sharing from the Network to the UE for Two-Sided Model Training Qualcomm Incorporated discussion Rel-19

[14] R2-2500614 Analysis on dataset and parameter transfer for two-sided model Lenovo discussion Rel-19

[15] R2-2500836 On Evaluation of Standardized Signaling for Two-side model ZTE Corporation discussion Rel-19 NR\_AIML\_air-Core

[16] R2-2500910 Discussion on signalling feasibility of dataset and parameter sharing for CSI compression Samsung R&D Institute UK discussion

[17] R2-2500998 Discussion on RAN1 LS on Dataset and Parameter Transfer Nokia discussion Rel-19 NR\_AIML\_air-Core