**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 20th, 2025, 10:00UTC.

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

|  |  |  |
| --- | --- | --- |
| **Company** | **Name** | **Email Address** |
| ZTE | Fei Dong | Dong.fei@zte.com.cn |
| Apple | Peng Cheng | pcheng24@apple.com |
| Huawei, HiSilicon | Jun Chen | jun.chen@huawei.com |
| vivo | Boubacar Kimba | kimba@vivo.com |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# 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] |
| ZTE | A1: No  A2: No  A3: Yes  A4: No  A5: Yes From RAN2 perspective | A1: I understand that only the average value can be considered as a requirement. it is so odd and not clear why only the minimum size is mentioned but the maximum size is not. It is suggested that :  -A1: In average, dataset and/or parameter sharing size can be as large as 225MB+11.6MB;  A2: In my understanding, the requirement of continuity is regardless of the duration or delay of data set and/or parameter sharing since there is no need to force one UE to receive the whole model parameters and/or data set, furthermore, one UE is not likely to stay in the RRC Connected state for days , or even weeks. So, it makes more sense to have the requirement of continuity is only for overcoming the data interruption due to the mobility. In addition, the continuity is not only for model transfer/delivery, but also for the data set sharing, so we suggest to have the following modification:  - A2: The continuity of model transfer/delivery and/or data sharing needs to be supported during the mobility. ;  A4: It is not clear about the meaning of user traffic, we can make it clear with 3GPP style wording  - A4: Low priority/QoS than CP/UP data transmission, with relaxed latency requirement and infrequent update; |
| Apple | A1: No  A2: No  A3: Yes  A4: No  A5: No | **A1:**  We share same view as ZTE. It does not make sense to consider minimum size due to below reasons:   1. It will imply that two model transfer solutions may be needed (one for small size and one for large size). RAN2 should target for a unified solution. 2. It may imply RAN2 prefer option 3a-1 without target CSI because 36kB is just minimum size of option 3a-1 without target CSI. 3. This is requirement and mentioning average value is sufficient.   Thus, we suggest below change:  -A1: In average, dataset and/or parameter sharing size can be as large as 225MB+11.6MB;  **A2:**   1. We share same view as ZTE. It is quite odd to couple service continuity with transfer duration (i.e. days/weeks). Because service continuity is RAN2 expertise, we don’t think any extra and unnecessary explanation is needed. 2. We are not sure why only model transfer is mentioned. It should be “dataset and/or parameter transfer / delivery” according to RAN1 LS.   Thus, we suggest below change:   * A2: ~~Model transfer/delivery~~ Service continuity of dataset and/or parameter transfer/delivery needs to be supported ~~considering dataset and/or parameter sharing may be expected to transfer in days/weeks~~;   **A4:**  Please note that priority/QoS are solution rather than requirement. As we are discussing requirement in this Question, we believe the 2nd half is sufficient:   * A4: ~~Low priority/QoS than user traffic, with~~ relaxed latency requirement and infrequent update;   **A5:**  We think it is not “model” but “dataset and parameter” have the requirement of open format and known structure because the UE needs to perform re-training (i.e. first training a nominal decoder and then training the actual encoder against the nominal decoder) based on the received “dataset and parameter” from NW.  Thus, we suggest below change:   * A5: ~~Model~~ visibility: open format and known structure for dataset and parameter. |
| Huawei, HiSilicon | A1-A4: Yes  A5: with comments | We suggest to keep the orginal text of A1 due to the following reasons:  (1) The value 36KB is clearly mentioned in the RAN1 LS, and here is the relevant text below. If companies think this value does not make sense, we would like to understand more about the reasons.  **For Option 3a-1 without target CSI: sharing encoder parameter**:   * + The number of parameters in the encoder models used in RAN1 evaluations have a wide range, ranging from 36K parameters to 13M parameters, with the median value of 5.8M.   + Based on the above range, the size of encoder parameters may range from 36K\*(8/8)=**36KB** based on the smallest model size in RAN1 evaluations and the use of float8, to 13M\*(32/8)=5**2M** based on the largest model size in RAN1 evaluations and the use of float32. Based on the median size, RAN2 may assume the size of 5.8M\*(16/8) = **11.6 MB** based on the medium model size and the use of float16.   (2) We are in phase 1, and we are just discussing requirements of model parameter and/or dataset (e.g. size, latency). We think we should follow what RAN1 have provided rather than doing some "filtering" as they are leading this discussion.  For A5, it is about visibility, and RAN2 ageed on the following for UE-sided data collection discussion:   * Note 2: Visibility of data content signifies that the MNO can, at least, be aware of, access, and comprehend the data without the need of SLA.   So we have the following suggestion on A5:   * A5: MNO visibility of data content of dataset/parameter   In addition, we think this email discussion is only about dataset/parameter, as mentioned in the RAN1 LS R2-2500015, and the direction is from NW to OTT server (with/without involving Uu interface). Others are out of scope, e.g. model structure, information transferred from OTT server to UE. |
| vivo | A1: Yes, with comments  A2: No  A3: Yes  A4: No  A5: See comments | **A1**: The total size of Option 3a-1 with target CSI is not the sum of Option 4-1 and 3a-1 without target CSI, since the dataset only contains the target CSI but not the CSI feedback, assuming N2=N1, the typical size of dataset and model parameter is 161.6 MB. In addition, agree with HW to indicate both minimum and medium size included in the R1 LS.  **A2**: the original A2 of model transfer/delivery continuity in the TR is considered due to UE mobility. However, the days/weeks is the latency of overall dataset transfer. Therefore, prefer revise it as:   * **A2**: ~~Model transfer/delivery continuity needs to be supported considering dataset and/or parameter sharing may be expected to transfer in days/weeks~~ Dataset/parameter sharing latency: relaxed latency (e.g., days/weeks) and be infrequent.   **A4**: Agree with Apple that A4 is solution rather than requirement, can be refined as:   * **A4**: Low priority/QoS than user traffic, with relaxed latency requirement and infrequent update Limited/no impact on the existing interaction over the air, including both signalling and user traffic.   **A5**: For the parameter sharing in Option 3a, the model format should be known to the UE side. Thus, it is not a requirement to be considered in R2 discussion, but is a precondition of Option 3a. |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

##### 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** |
| ZTE | The feasibility of A5shall be confirmed by RAN1. |
| Apple | Response to ZTE: A5 is a necessary requirement. The UE needs to perform model re-training based on the received “dataset and parameter” from NW. Thus, the UE needs to be able to decode and fully comprehend the dataset and parameter” from NW. Otherwise, the two-sided model can’t work. We think it is RAN1 common understanding, and no need to confirm with them. |
| vivo | Agree with Apple. |
|  |  |

## 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)

|  |
| --- |
| base station, cell tower, communication, connection, network, tower cloud, database, hosting, server cloud, server, web  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)

|  |
| --- |
| 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 |

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)** |
| ZTE | No for Alt.1  No for Alt.2 | For alt.1, we do not think the data path from gNB to the NW dataset/model parameters collection entity, and from UE server to the UE is in the scope of this email discussion, the data path for alt.1 is as following:  NW dataset/model parameters collection entity -> UE training entity  For alt.2, in RAN1 LS, the OTA approach means the NW will share the dataset or model parameter with UE via the air-interface. And this email discussion is only to focus on model transfer/delivery and data set sharing from NW to the UE, then the dataset collected by gNB sending to the NW dataset/model parameters collection entity as well as the data path from UE to UE server is **NOT** in this email discussion scope, and hence the data path shall be as following:  gNB -> UE |
| Apple | No for Alt-1 and Alt-2,  And **RAN2 should focus on issues in RAN1 LS** | 1) Although we appreciate Rapporteur’s efforts on solution analysis in the table, we have to remind that **RAN1 only ask RAN2 on option 4 and option 3a-1 in** their LS (R2-2500015). We understand that Rapporteur is ambitious to consider all options (e.g. option 3b) in RAN2. However, please note according to SID, **RAN2 is NOT involved in objective of CSI-compression**:    Thus, business as usual, RAN2 is not necessary to overdo task out of scope and only need to study the issue raised by RAN1 LS**.**  **As summary, we suggest RAN2 only study issues of two-sided model raised in RAN1 LS (i.e. only option 4 and option 3a-1 is in scope).**  2) Following 1), we agree with ZTE that RAN2 study on OTA and non-OTA approach should only focus on issues raised in RAN1 LS, i.e. we suggest below change:  **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)~~  **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)~~  Then, on the figure of non-OTA and OTA, we appreciate Rapporteur’s efforts to align company E2E understanding but also understand QC’s concern on potential being out of scope.  **As compromise, we suggest keeping the two figures, but:**   * **Highlight the line between NW dataset/model parameters collection entity and UE training entity in non-OTA figure** * **Highlight the line between gNB and UE in OTA figure.** * **Add a note that other procedures / interfaces can be studied in normative phase (if any).** |
| Huawei, HiSilicon | Yes, with comments | Regarding the table, we have some comments/suggestions:  (1) For Option 1&2, requirement of offline training is empty. For option 1, RAN1 has made the following agreement, and we think RAN1 may discuss it later.   * Specification of option 1, if needed from RAN1, can reuse specification of option 3a/3b, with the additional specification of parameters.   It may cause some confusions, e.g. Option 1&2 are relevant to on-device operation without offline engineering. In this case, we sugges the following:  for option 1&2, the requirement of offline training: pending for RAN1 progress  (2) for Option 3b (z4), there is a text from RAN1 LS (below).   * Direction B: Sharing NW side encoder parameter to UE side for UE side inference directly with on-device operation (Inter vendor collaboration option 3b), including at least the following issues   and then the definition is also shown below. For now, the information for option 3b is empty in the above table, so it may cause some confusions, e.g. whether model structure+parameters or only parameters are exchanged for option 3b.  We suggest to add: CSI generation part   * + Option 3b     - The method of exchanging is over the air-interface via model transfer/delivery Case z4.     - The parameter exchange is from NW to UE.     - Parameters exchanged from the NW-side to UE-side is CSI generation part.   In figure for Alt 1, we suggest to remove the last step, i.e. from OTT server to UE, as it is out of RAN1 LS's scope.  In figure for Alt 2, similar comments here, i.e. the model transfer from OTT server to UE should be removed. In addition, for the 2nd entity, the text below "dataset transfer" can be removed as it is a bit confusing. |
| vivo | Yes, for Alt1;  No for Alt2 | For **Alt 1**, fine with the intention of moderator to illustrate the overall procedure.  For **Alt 2**, the data transfer path is confusing, the data is always from the gNB to UE for the OTA approach. Thus, suggest refining as follows:  **gNB** -> **NW dataset/model parameters collection entity -> gNB** -> **UE** -> **UE training entity**  Note that the data path for these two approaches are only for alignment of understanding. Agree with ZTE and apple that R2 only focus on:  Alt 1: dataset from collection entity to training entity;  Alt 2: gNB to UE. |
|  |  |  |
|  |  |  |

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** |
| ZTE | Yes | Please see our comments in Q 2-0. |
| Apple | Partially Yes, but it is out of scope of email discussion. | See our comments in Q2-0. It is out of RAN2 scope. And because it has no RAN2 impact, we don’t see any emergency for RAN2 to study it in Rel-19. It can be totally postponed to normative phase (if any).  If majority prefer to conclude it in RAN2, we are fine but RAN3 should be added because it is interface between gNB and OAM/CN. |
| Huawei, HiSilicon | Yes | Firstly, we are open to study it in RAN2, like we did for UE-side data collection in the past.  Secondly, regarding details and feasibility, we do not think RAN2 can progress on them.  As a result, we think RAN2 could try to provide some analysis just from RAN2 point of view. |
| vivo | Yes, with comments | RAN3 is also involved. |
|  |  |  |
|  |  |  |

### 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 (gNB/OAM/CN/gNB side server) 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. OAM -> UE-side OTT server | SA5 | Up to SA5 |
| 1. 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] |
| ZTE | 1. Yes 2. Yes 3. No 4. No | Option 3 and Option 4):From NW vendor point of view, they are not allowed due to the security and privacy issue that is caused by directly transferring the data to the outside, it shall be removed from the feasible options. |
| Apple | 1. Yes 2. Yes 3. Yes 4. No | For 1) and 2), we think no technique showstopper from RAN2 perspective, but technique details can be left to SA2/SA5/RAN3. Thus, we prefer:   * Conclude RAN2 assume the following options are feasible:   + OAM -> UE-side OTT server   + CN -> UE-side OTT server. * Send LS to SA2/SA5/RAN3 to check any issue.   For 3), we think it should be feasible in case of offline engineering. But RAN2 may not need to study it.  For 4), we are confused why two servers are needed, and questioned it is unnecessarily making things complex. |
| Huawei, HiSilicon | 1), 2), 3): Yes | For 4), we are unclear what is the difference between 4) and 2). For UE-side data collection, RANP has decided to exclude option 1b, and option 1b is similar to 4) here.  So we think the necessity of 4) (on top of 1)) should be clarified first.    For 3), we are not sure whether it is outside 3GPP and up to NW implementation. Maybe SA5 can take a look at it.  Regarding the wording, we see there are different definitions of "server", and we are a bit confused. The wording "UE-side OTT server" is ok as it is anyway outside 3GPP networks. We suggest to align the wording, e.g.  1) gNB -> OAM -> UE-side OTT server  2) gNB -> CN -> UE-side OTT server  3) gNB -> UE-side OTT server ~~(outside of MNO)~~  4) gNB -> server inside MNO -> UE-side OTT server ~~optionally OTT server (outside of MNO)~~ |
| vivo | 1. Yes 2. Yes, with comments 3. No   4) No | For 2), gNB->CN is in R3 scope as comment in Q2-1, while CN-> server is out of R3 scope.  3) and 4) may have impact on R3 specification. |

##### 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** |
|  |  |  |  |
|  |  |  |  |











### 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.

##### 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** |
| ZTE | Solution 1a Yes  Solution 1b  No | Technically for option 1a, the RRC signaling between gNB and UE is a traditional solution to transfer the data between UE and gNB, it definitely can be reused for OTA approach. But the super size of the data set /model parameter may bring the challenge to the current size limitation of RRC signaling, how to overcome such challenge can be further discussed.  For option 1b, we do not think this is a feasible option in NR stage since there is no UP tunnel terminated between UE and gNB so far. It can be excluded from the feasible solutions. |
| Apple | Solution 1a Yes  Solution 1b  No | Same view as ZTE.  For option 1b, we think it has too much spec work to support a new protocol stack. If any interest, it should be proposed in 6G. |
| Huawei, HiSilicon | Yes | In our Tdoc R2-2501111, we also proposed to list them as candidate options. Here we would like to clarify two things:  (1) we think the initiating node should be discussed first, and we assume that NG-RAN can generate dataset/parameter. In other words, if other nodes would like transfer these information to UE, there will be co-ordination between NG-RAN and these nodes, so we also need to take these aspects into account  (2) in the current TR 38.843, we have had some analysis for model transfer/delivery options, and Solution 1a/1b in TR 38.843 are quite similar as Solution 1a/1b here. So we think the existing analysis can be re-used as much as possible |
| vivo | Solution 1a Yes  Solution 1b  No | Agree with ZTE. |

##### 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** |
|  |  |  |  |
|  |  |  |  |



#### 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] |
| ZTE | 1. Yes 2. Yes 3. Yes |  |
| Apple | 1. Yes 2. No 3. No | We support 1).  For 2) and 3), we have some confusion about their difference from 1). We provide our view based on the understanding that gNB doesn’t need to decode and comprehend dataset/parameter before forwarding to UE (otherwise, 2/3 are same as 1). In detail:   * 2) means CN ->UE via gNB transparently relay (e.g. similar to NAS) * 3) means OAM->UE via gNB transparently rely   With above understanding, we doubt whether 2) and 3) with the identified solution (solution 2a/2b for 2 and solution 4b for 3) are necessary. The solution 2a/2b for 2) and Solution 4b for 3) were identified for model transfer / delivery. We don’t think RAN2 can conclude they can be reused to “dataset transfer / delivery” because of the below difference:   * In Model transfer / delivery: one single UE needs to get complete model parameter and model structure. Thus, gNB doesn’t need to comprehend the model and thereby the direct path (e.g. NAS signaling in solution 2a/2b) makes sense. * In dataset transfer / delivery: one single UE doesn’t need to get complete dataset but can get just a split part for training. Thus, in case of dataset splitting, gNB needs to comprehend the content of dataset and thereby the direct path (e.g. NAS signaling in solution 2a/2b) doesn’t make sense because gNB is transparent to content of NAS signaling. |
| Huawei, HiSilicon | 1): Yes  2)&3): benefits over non-OTA are unclear | In our Tdoc R2-2501111, we also provided some analysis regarding 2) and 3), and here are the key parts:  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  **Proposal 2: For over-the-air approach, it is assumed that NG-RAN generates the dataset/parameter.**  Once NG-RAN has generated dataset/parameter, it may either directly send them to UE via Uu interface, or send them to CN/OAM. For the later way, CN/OAM can further transfer them to OTT server, i.e. other approaches. If CN/OAM transfer them to NG-RAN, and then to UE, the transmission path would be: NG-RAN -> CN/OAM -> NG-RAN -> UE -> OTT server, and we do not think it has benefits over other approaches (detailed analysis can be found in section 2.3). So the above solution 2a/3a/2b/3b/4b are not considered for analysing the over-the-air approach. For the above solution 4a, it is about the model transfer between OTT server and UE, and it is not related to this over-the-air approach. In summary, we think the above solution 1a and 1b can be considered for the over-the-air approach.  \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  For the highlighted part, we compared "gNB->CN/OAM->OTT server" with "gNB->CN/OAM->UE->OTT server", and we think the 1st path has more benefits. We are open to this comparison, and maybe companies can figure out more benefits for the 2nd path |
| vivo | 1. Yes 2. Yes 3. Yes | Options 2) and Option 3) can be beneficial for centralized Model training, and the model can be applicable across multiple gNBs. |

##### 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** |
| ZTE | No, just option 1a | In our understanding, NW transfers the data which has been desensitized to the UE for UE to train the AI/ML model, there is no need for NW to be aware of or even get involved the data transfer between UE and its OTT server. |
| Apple | Agree option 1a,  We are fine to send LS to SA2/SA5 on option 2 and 3 | First, we think it is straight forward to agree option 1a can work.  Then, we understand that Rapporteur seems to imply that UE-side data collection solution is reused for two-sided model. However, we don’t think RAN2 is in position to make this conclusion due to below reasons:   * It is common understanding that SA2 has last call for UE-side option 2 and SA5 has last call for UE-side option 3. Then, we are confused why RAN2 can make conclusion to reuse same solution for two-sided model without checking SA2 and SA5? * From technical perspective: we also don’t think the same solution can be reused:   + In two-sided model, NW share its dataset / parameters in open format to UE. Then, we are not sure why NW needs to be aware of the UE transfer procedure for the dataset/parameters which are generated by itself.   + In our understanding, the requirements for UE-side data collection and two-side model are different. Take privacy as example:     - In UE-side data collection, it is UE to share its dataset with NW but no requirement for the NW to send its dataset to the UE. Thus, the privacy concern is only UE privacy concern.     - In two-sided model, NW needs to share its dataset with UE. Thus, it has NW privacy concern.   Based on above analysis, we believe that RAN2 can send LS to SA2/SA5 to ask them to decide whether option 2 and option 3 of UE-sided data collection can be reused to two-sided model. |
| Huawei, HiSilicon | OK with 1a  Comments on option 2/3 | Firstly, for UE-side data collection, we think only Option 1a is supported in Rel-19, and RAN2 also made some progress regarding data collection configuration, e.g. the network can provide the data collection configuration (at any point in time), with or without UE request. For option 2 and 3, they will be studied until the end of Rel-19.  Secondly, as we commented online, we think that with OTA approach, how UE is to transfer dataset/parameter can be let to implementation, i.e. option 1a here. If UE-OTT server is to use option 2/3, the transmission path will be:  NG-RAN -> (other NW nodes) -> UE -> NG-RAN -> CN (option 2)/OAM (option 3) -> OTT server  It means that NG-RAN sends the information to the UE, and then the NG-RAN receives the exactly same information from the UE, which is very strange to us.  In summary, for OTA approach, we are open to discuss the above solutions for UE-side data collection, but we are unclear about the necessity of option 2/3 (as RANP has excluded 1b here. |
| vivo | 1a only | Oher cases are not valid, i.e., if CN or OAM can transfer data to the training entity, there is no need for the UE to relay the dataset. |

# 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