**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 |
| Qualcomm | Rajeev Kumar | rkum@qti.qualcomm.com |
| Lenovo | Congchi Zhang, Tapisha Soni | zhangcc16@lenovo.com, tsoni@lenovo.com |
| Samsung | Milos Tesanovic | m.tesanovic@samsung.com |
| OPPO | Jiangsheng Fan | fanjiangsheng@oppo.com |
| Nokia | Jerediah Fevold | jerediah.fevold@nokia.com |
| LGE | Minji Choi | stacyminji.choi@lge.com |
| CMCC | Ningyu Chen | chenningyu@chinamobile.com |
| Futurewei | Chunhui (Allan) Zhu | czhu@futurewei.com |
| Ericsson | Marco Belleschi | marco.belleschi@ericsson.com |
| CATT | Tangxun | tangxun@catt.cn |

# 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. |
| Qualcomm | A1: Yes (with modification)  A2: Yes (with modification )  A3: No  A4: Yes (With Modification)  A5: No | **A1:**  We share the same view as ZTE and Apple. Additionally, we want to highlight that RAN1 LS explicitly mentioned that for the “Option 3a-1 without target CSI: sharing encoder parameter” as below:   * 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.   Therefore, it is not reasonable for RAN2 to consider the minimum size of it, when RAN1 explicitly asked RAN2 to consider median size as 11.6 MB for RAN2 study.  **A2**:  We share the same view as ZTE and Apple. We support Apple's modification for A2.  **A3:**  We believe that the dataset and parameters originate from the NW, therefore, network control on sharing of dataset/parameters is obvious. Maybe the rapporteur should further clarify if network controllability implies here “transfer/delivery of dataset/ parameters from UE to OTT server” once UE receives the dataset/parameters (in case of OTA sharing of dataset/parameters).  **A4:**  We agree with Apple update.  **A5:**  The dataset and parameters are expected to be used by the UE side for training. Therefore, we believe that the dataset and parameters are in a format that is understandable by both UE and NW. |
| Lenovo | A1: no, with comment  A2: yes with modification  A3: yes  A4: yes with modification  A5: yes, with modification | One general comment as raised by some companies, it would be more clear to separate the dataset/parameter transfer from the model delivery. They are two different steps.  A1: **We understand the data size requirement depends on which option of 4-1, 3a-1, 3a-1 with target CSI will be supported at the end**. From RAN2 study point of view, we can analyse the feasibility for each option 4-1, 3a-1, 3a-1 with target CSI. If we want to have one single requirement covering all cases, then we need to take the worst case scenario 255MB + 11.6MB.  A2: ~~Model transfer/delivery~~ **Dataset/Parameter transfer continuity** needs to be supported considering dataset and/or parameter sharing may be expected to transfer in days/weeks.  A4: This seems only applies to OTA based approach. Besides, low priority/QoS than user traffic implies e.g., RB with lower priority than other SRB/DRB, which is rather stage 3 details. Maybe we can just repeat the ran1 statement:   * A4: ~~Low priority/QoS than user traffic, with~~ relaxed latency requirement and infrequent update;   A5: ~~Model~~ **Dataset/Parameter** visibility: open format and known structure. |
| Mediatek | A1-A4: Yes with comment  A5: No | **A1:**  To avoid confusion, the requirements for model parameters and dataset transfer pertain to their size, which can range from tens of KBs to hundreds of MBs. For UE implementation, it is important to consider not only the minimum size but also the maximum and typical sizes.   * A1: ~~Minimum~~ The dataset and/or parameter sharing size can be ~~36kB~~ ranging from tens of KBs to hundreds of MBs. In average, dataset and/or parameter sharing size can be as large as 225MB+11.6MB;   **A2:**  A2 mixed the requirements for model transfer/delivery and model parameter/dataset sharing for model training. According to the RAN1 LS, it is clear that since the purpose of dataset and/or parameter sharing is for UE-side offline training, this process is expected to have relaxed latency (e.g., days/weeks) and be infrequent. However, this does not necessarily imply that model transfer/delivery has the same requirements. For example, if a cell-specific two-sided model is developed, it is possible to have reactive model transfer when the UE moves to the corresponding cell. (Note: I am not proposing this approach but just using it as an example.) Similar to ZTE's suggestion for A2:   * A2: The dataset and/or parameter sharing continuity during UE mobility needs to be supported.   **A3:**  Not sure about the intention of ‘securely’ in the description. ‘Securely’ is the natural requirement for the system to delivery any data over CP or UP.  **A4:** Priority is not equivalent to QoS, which considers more aspects and QoS mechanisms use priority levels to manage and control the traffic flow. To be percisesly:   * A4: Low priority~~/QoS~~ than user traffic, with relaxed latency requirement and infrequent update   **A5:**  The concept of an open format and known structure is currently being studied in RAN1, but no conclusions have been reached yet. Regardless of whether we are discussing collaboration level z4 for model transfer or Direction C for model parameter/dataset sharing for UE-side offline training, RAN2 needs to wait for further progress from RAN1 on this aspect to avoid duplicated discussions and conflicting understandings. |
| Samsung | Yes – A1, A3  No – A2, A4, A5 | Regarding concerns from ZTE and Apple on A1, we do think the minimum data size is important as it determines whether e.g. segmentation would be needed (in case minimum size of the messages is above a certain value).  A4 is a network implementation matter.  A5 – for the specific case under discussion as captured in RAN1 LS, the data set format is already assumed to be open/known to both sides. We are ok with Apple’s revision for A5.  A2 – we do not think continuity (in the sense of sequential delivery) is an essential requirement as concatenation/segmentation could be handled by OTT server and/or gNB. Or are we talking here about continuity in case of e.g. mobility? We additionally do not understand why continuity is linked to transfer period duration. Also, we still need to ask RAN1 to clarify the ‘days/weeks’ range for A2 (this may not affect the protocol for data set and parameter transfer, although it could affect the choice between different options e.g. time range could result in high overhead if CP is used, depending on the time window for transfer). |
| OPPO | A1: No  A2: No  A3: Yes  A4: No  A5: No | A1: similar view as Apple and ZTE, considering normal/average scenario will be better as it will be more future friendly.   * A1: In average, dataset and/or parameter sharing size can be as large as 225MB+11.6MB;   A2: similar view as Apple and ZTE, as RAN1 LS uses the wording dataset and/or parameter sharing, so we suggest the following revision for better alignment .   * A2: Service continuity of dataset and/or parameter sharing needs to be supported;   A4: similar view as Apple   * A4: Relaxed latency requirement and infrequent update;   A5: As clarified in the very beginning, RAN1 LS focus on dataset and/or parameter sharing, model visibility is only applicable for model transfer, to better track RAN1 focus, we suggest the following re-wording:  Visibility for dataset and/or parameter sharing: open format. |
| Ericsson | A1: Yes with modification  A2: Yes with modifications  A3: Yes with modifications if needed  A4: OK with Apple proposal  A5: NO | A1: We are ok with the modifications proposed by Apple/ZTE/QC (and other above). The solution to be designed should take into account at least the average case (225MB+11.6MB)  A2: We are ok with several of the comments above suggesting to focus only on the UE mobility. We are ok with Mediatek or ZTE proposal above.  A3: In both OTA and non-OTA approaches, the NW controllability is present/inevitable, since in both cases the transfer of dataset/model parameters is initiated by the NW. If there are concerns on the “securely” (as pointed out by Mediatek), it can be rephrased like, since how to ensure security it will be implicit in the protocol used for the exchange of data, irrespective of whether it will be OTA or non-OTA delivery: “NW controllability: Decision on if and when to transfer/delivery the dataset and/or model parameter ~~securely in a NW-aware manner~~”  A4: We are ok with Apple proposal  A5. After checking RAN1 agreements, we suggest some rewording of this requirement, since this requirement was written with the discussions on the model transfer in mind. From RAN1#117, it is clear that something needs to be standardize to ensure visibility between NW and UE parties. However, it is not clear whether this would imply an open format or not. We therefore prefer the following rephrasing, in order to just focus on the need for 3GPP to standardize the format/structure (rather than on the type of format):   * A5: ~~Model~~ visibility of dataset/model parameters: ~~open~~ standardized format/ ~~and known~~ structure for the dataset/model parameters. |
| Nokia | A1: No (see comment)  A2: No  A3: Yes  A4: No (See comment)  A5: No (See comment) | **A1**: We have to support realistic dataset and parameter set sizes. Although we do not support this solution, we think that the minimum values provided by RAN1 should be adopted to understand the problem. Since both datasets and parameter sets are on the table, we should consider the minimum size to be 237MB.  **A2**: We do not agree to create a complex solution that requires mechanisms on top of those which already exist to facilitate retransmissions and continuity, and to detect corrupt transmissions.  **A4**: We agree with Apple and OPPO on reducing the description of A4 to simply include a relaxed latency requirement.  **A5**: The requirement for an open format and known structure should be applied to the end-to-end transfer, which is between the NW and the server for training UE-side models. There is no requirement on model visibility between the UE and gNB for this dataset and parameter set transfer. Therefore, A5 isn’t in our scope.  Generally, we do not think that the evaluation criteria we used for model transfer of a UE-side model to a UE are appropriate for the evaluation of using UEs to transfer datasets and parameter sets used for training UE-side models. Additionally, the Tdoc submitted by T-Mobile on requirements and most Tdocs submitted on this topic thus far were about model transfer/delivery to a UE, not about dataset and parameter set transfer from the NW to the UE-side server. |
| LGE | A1: Yes (with comments)  A2: Yes (with comments)  rA3: Yes  A4: Yes (with comments)  A5: No | A1 : We share the similar view with Mediatek that it is important to consider not only the minimum size but also the maximum and average sizes as a requirement.   * A1: ~~Minimum~~ The dataset and/or parameter sharing size can be ~~36kB~~ ranging from tens of KBs to hundreds of MBs. In average, dataset and/or parameter sharing size can be as large as 225MB+11.6MB;   A2 : We share the same view as Apple. The discussion of service continuity with transfer duration(i.e., days/weeks) is odd and not needed to be mentioned. It is important for RAN2 to also strictly adhere to the RAN1 LS regarding "dataset and/or parameter transfer/delivery" and not to selectively choose parts of its content, such as "model transfer/delivery," as this could lead to further confusion.   * 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: We share the same view with Apple,vivo, Lenovo and OPPO. We think that the low priority/QoS in A4 is solution rather than requirement and needs to be discussed in stage3.   * A4: ~~Low priority/QoS than user traffic, with~~ relaxed latency requirement and infrequent update;   A5 : We think the clarification is needed regarding what exactly is meant by "visibility". In our opinion, the term "visibility" here seems to refer more to whether the datasets and parameters transmitted to the UE are in an open format, rather than being something that the MNO can comprehend. Since the gNB transfers the dataset/parameters, MNOs would naturally have the visibility of data content of dataset/parameter. |
| CMCC | A1: No  A2: No  A3: Yes  A4: No  A5: Yes | A1: We have comments on the maximum dataset size, the total size are as follows based on RAN1 LS:   * For Option 4-1, the total dataset size may be roughly **225 MB**. * For Option 3a-1 without target CSI, the size may range from 36KB to **52MB**, and the medium size is 11.6MB. * For Option 3a-1 with target CSI, the size becomes N2 \* (2000 bits) / (8bits/Byte) + 11.6 MB, and N2 may be the same or smaller than N1 (600K). The maximum size may be 600K \* (2000 bits) / (8bits/Byte) + 11.6 MB ≈ **158.1 MB**, when N2 is the same as N1.   Based on the above analysis, the maximum size should be 225MB. And we share similar view as ZTE and Apple, so it is suggested to update A1 as:  -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: We think A2 mixed the requirements for service continuity during UE mobility and latency of dataset and/or parameter sharing. Considering A4 is requirement on latency, we prefer MTK’s suggestion for A2.  A4: Agree with Apple’s modification.  A5: We are also fine to use “dataset and/or parameter” to replace “model”.  From the perspective of operator, A3 and A5 are necessary to guarantee the security and privacy. |
| Futurewei | A1: No  A2: No  A3: Yes  A4: No  A5: No | A1: for standard/protocol design point of view, median/average size and maximum size (the worst case) are more important than the minimum size. As a compromise, we can list three sizes, minimum, median, and maximum.  A2: agree with many companies above that continuity should primarily address UE mobility and data interruption issues rather than being linked to the duration or delay of dataset/parameter sharing.  A4: we are not sure how to use a relaxed requirement to evaluate a solution; for a relaxed requirement that all or most solutions can meet, the requirement/criterium is not useful as it can be easily met.  A5: we believe the question itself needs clarification. If it was intended to mean “model visibility” then we would argue that the issue does not exist; the visibility is between the NW and the UE, and it should always be visible. We suggest clarifying the question so companies can provide their opinions again. |

##### Summary:

|  |  |  |
| --- | --- | --- |
| **Discussion Area/Requirement** | **Proponent** | **Opponent** |
| A1: Dataset/model parameter size | HW, vivo, QC (but remove minimum size), MTK, SS, Ericsson, LG | ZTE, Apple, Lenovo, Oppo, Nokia, CMCC, FW  Reason:  1. Do not consider minimum size, and consider a unified solution for both small/large size (ZTE, Apple, Lenovo, QC, CMCC)  2. minimum size to be 237MB (Nokia) |
| A2: Continuity | HW, QC (with Apple’s update), Lenovo (dataset/model parameter), MTK, Ericsson, LG (with Apple’s update) | Apple (with updates), ZTE, vivo, SS, Oppo, MCC, FW  Reason:  1. UE doesn’t need to receive the whole model parameter/dataset (ZTE)  2. not coupling with latency requirement (Apple, vivo, SS, CMCC)  3. no need for complex solution on top of existing mechanisms for retransmission/continuity/detect corrupt transmission (Nokia) |
| A3: NW controllability | ZTE, Apple, HW, vivo, Lenovo, MTK (not sure about securely as it’s natural), SS, Oppo, Ericsson, QC (from NW-> UE is yes), Nokia, LG, CMCC, FW |  |
| A4: Priority/QoS | HW, QC (with Apple’s update), Lenovo (only latency), MTK, ZTE (with clarification), LG (only for latency requirement) | Apple (with updates), vivo, SS, Oppo, Nokia, CMCC, FW  Reason:  1. use latency requirement rather than priority/QoS (Apple, Vivo, Lenovo, CMCC)  2. priority is an implementation matter (SS) |
| A5: Visibility | ZTE, HW (MNO visibility), Lenovo, Apple (on dataset/model parameter only), CMCC (on dataset/model parameter only) | vivo, QC, SS, Oppo, MTK, Ericsson, LG, FW  Reason:  1. it’s a precondition not a requirement (vivo)  2. wait for RAN1 to avoid duplicate discussion (MTK)  3. only focus on open format (Oppo)  4. focus on dataset/model parameter only (Ericsson, LG)  5. no requirement on visibility between UE and NW (Nokia) |

For A1, it is also true that RAN1 LS highlighted several sizes of dataset/model parameter in bold:

* Option 4-1: 225MB
* Option 3a-1 without target CSI: 36KB smallest, 52M largest, RAN1 further suggest RAN2 to assume 11.6MB as medium model size with the use of float16
* Option 3a-1 with target CSI: N2 \* (2000 bits) / (8bits/Byte) + 11.6 MB (not equal to 225MB+11.6MB)

However, rapporteurs understand that RAN2 should aim to define a unified solution to support various size of dataset/model parameter.

For A2, rapporteurs further clarify it is for continuity during UE mobility. It is not the intention from Rapporteurs to indicate any new mechanism is needed for service continuity. Existing solutions to support continuity during UE mobility can be reused if suitable.

For A3, all companies agree that NW needs controllability for dataset/model parameter transfer from NW to UE-side OTT server. One company question whether this is the requirement for the link between UE to UE-side OTT server. Based on rapporteurs’ understanding, controllability/requirement for the link from UE to OTT server is the same as UE-side data collection. Hence, we don’t need to duplicate the discussion here.

For A4, more than half of companies would like to focus on the latency requirement only. Rapporteurs update A4 as latency requirement and the updates can be found in below.

For A5, most of companies prefers to focus on dataset/model parameter only. One company mentioned the format needs to be understandable by both UE and NW. Another company mentioned there’s no requirement on visibility between UE and gNB. However, Based on RAN1 LS, no matter OTA/non-OTA method, they are both within the scope of Direction A, where standardized data format and model format is a general requirement for both solutions.

In the end, rapporteurs further align the wording with RAN1 discussion by using ‘standardized data format and model structure’.

Based on above summary, rapporteurs would like to propose following principles for dataset/model parameter transfer solution selection:

Proposal 1: Dataset/model parameter transfer solution from NW to UE-side OTT server/UE training entity shall follow below principles:

* **A1 -** **Size**: From RAN2 point of view, RAN2 aims to define a unified solution to support various sizes of dataset/model parameter transfer (dataset and/or parameter sharing size could range from tens of KBs to hundreds of MBs, but in average around hundreds of MBs);
* **A2 - Continuity**: Service continuity of dataset and/or parameter transfer/delivery during UE mobility needs to be supported;
* **A3 - Controllability**: NW controllability: Decision on if and when to transfer/delivery the dataset and/or model parameter from NW to UE training entity (OTT server inside/outside of MNO);
* **A4 - Latency**: Relaxed latency requirement and infrequent update;
* **A5 - Visibility**: standardized data format and model structure for dataset and parameter to be understandable by UE.

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

|  |  |
| --- | --- |
| **Company** | **Comment** |
| Qualcomm | For the dataset/parameter sharing, RAN2 should further consider the below aspects:   * Impact on Uu resources (due to size of dataset / parameters) * UE impacts (like power impact, memory requirements, etc), * UE selection (how UE selection for sharing of dataset/parameters are performed), * End-to-end reliability of dataset/parameter sharing (how to ensure that a UE vendor receives the required dataset/parameters). |
| Samsung | Consideration of legal framework should be done e.g. whether server-to-server transfer is possible between NW and UE OTT servers. Additionally, the network topology or any NW vendor proprietary information should not be disclosed. |
| Ericsson | We are ok to consider the aspects listed by Qualcomm. At least during the discussions of the pros, and cons for the various solutions, those aspects should be taken into account by RAN2. |
| Nokia | The scope of this email discussion is limited to the transfer of datasets and parameter sets from the NW to the server for training UE-side models for the purpose of training UE-side models. We should not link this topic with model transfer/delivery as they are different and serve different purposes. |

##### Summary:

There are some additional principles proposed by companies:

* Uu interface impact (e.g. radio resource)
* UE impact (e.g. power/ memory requirement)
* UE selection for being reception of dataset/model parameter
* E2E reliability of dataset/model parameter transfer
* proprietary information of NW topology and NW vendor should be respected and not disclosed.

Due to limited response during phase 1, rapporteurs suggest to further discuss in phase 2 whether above additional principles should be added on top of Proposal 1.

##### 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. |
| Lenovo | Agree with Apple. |
| Mediatek | I believe we have conflated model transfer with model parameter/dataset sharing, as well as visibility with open format/known structure in this discussion. For model parameter and/or dataset sharing, it is certain that this information is visible to both the network side and the UE side. However, for model transfer—when the UE-side model is trained and transferred to the UE—it is unclear whether the UE-side model needs to be visible or defined with an open format/known structure. This matter is still pending further discussion in RAN1. |
| Samsung | RAN2 to ask RAN1 to confirm that the path for dataset and/or parameter sharing for non-OTA approaches is NW server to UE server (and without any 3GPP-standardised involvement of the UE), and to ask RAN1 to clarify what RAN1 believe RAN2’s remit here would be (if any).  RAN2 to ask RAN1 whether they envisage gNB being involved in managing the data set transfer.  RAN2 to ask RAN1 about their assumption on how the segmentation/concatenation of messages within a single data set is to be done in UE server and whether assistance from the 3GPP network is needed for this, and to further ask for the minimum message size anticipated. |
| OPPO | Agree with Apple, open format should be the baseline for cross-vendor collaboration solution. |
| Nokia | We agree with Mediatek and Samsung. The purpose of the transfer of datasets in the context of this discussion is not to provide a UE with a model to run, but rather to provide a server which trains UE-side models with training inputs. Therefore, it is not important for the UE to be able to understand the model structure. At which level beyond RAN2 scope the model format and structure are known in the parameter set and dataset transfer being discussed here isn’t up to RAN2. |

##### Summary:

Rapporteurs understand companies’ concern on ‘dataset/model parameter transfer’ vs. ‘model transfer’. Though in rapporteurs’ understanding, these two procedures are concatenated together, i.e. first NW transfers dataset/model parameter transfer to UE-side OTT server, then model transfer from UE-side OTT server to UE. Considering RAN1 LS only asks about NW -> UE OTT server for dataset/model parameter transfer, rapporteurs suggest we put aside ‘model transfer’ from UE-side OTT server to UE that were discussed during Rel-18.

Only two companies raised questions would like to further check with RAN1, the proposed questions are summarized as below:

1. Feasibility of A5 – Rapporteurs understand this is clear in RAN1 LS, and no need to ask.

2. Ask RAN1 to confirm path for non-OTA is NW server to UE server without 3GPP impact

3. gNB involvement in dataset transfer

4. segmentation/concatenation of messages within a single dataset to be done in UE server or inside 3GPP

Rapporteur thinks above question 2-4 is the scope of RAN2 discussion, we can conclude in RAN2 first whether this is a feasible solution and then reply to RAN1.

Observation 1: So far, RAN2 didn’t observe any new question/further clarification needed from RAN1 for feasibility analysis of dataset/model parameter transfer.

## 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.   **Rapporteurs’ response:**  Let’s focus on Direction A in this email discussion as asked by RAN1. The above table is just for information to RAN2 companies on the discussion background in RAN1.  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. |
| Qualcomm | Alt1: Yes (with modification)  Alt2: No (needs modification) | To Huawei: We believe this email discussion should focus on what RAN1 has asked RAN2 to study and provide feedback. RAN2 does not need to discuss aspects that is still under RAN1 discussion.  To VIVO: I believe that RAN2 discussion should be focused on dataset / parameters sharing after network-part of two-sided model training (that is outcome of network side training), and need is shared to support UE side model training . That data needed for NW-side training is discussed under NW-side data collection.  We appreciate the rapporteurs' update on ALT 1 to remove model transfer/delivery aspects from the email discussion and figures. We also want to highlight that based on RAN2 agreements during the SI, the network part of the two-sided model training may happen at the gNB, OAM, and CN. Therefore, the origin of the dataset/parameters may not even be gNB (at least in the cases where training happens at CN / OAM). Therefore, we suggest the following modification:  **For the non-OTA solution**: We agree with Apple to focus our discussion on data sharing from “**NW dataset/model parameters collection entity** -> **UE training entity”.**  Otherwise, RAN2 may need to separate the discussion into two sub-alternatives, as dataset/parameter sharing between gNB and NW dataset/model parameters collection entity is needed only if the network-part of two-sided model training is performed at gNB.   * **gNB** -> **NW dataset/model parameters collection entity** -> **UE training entity** (OTT server inside/outside of MNO) [if the network-part of two-sided model training is performed at NG-RAN] * **NW dataset/model parameters collection entity** -> **UE training entity** (OTT server inside/outside of MNO) [if the network part of two-sided model training is performed at OAM/CN]     **For the OTA solution:** We disagree with the rapporteur on the data transfer path for the OTA approach. There may exist two sub-alternatives.   * **gNB** -> **UE** [if the network-part of two-sided model training is performed at NG-RAN] * **NW dataset/model parameters collection entity -> gNB -> UE** [if the network-part of two-sided model training is performed at OAM/CN]     We want to highlight that the OTA approach may require solutions for the transfer of data from OAM / CN to the gNB if the training is not performed at the gNB.  Considering above, we believe that procedure is required to share the dataset / model parameter between gNB and NW dataset/model parameters collection entity, irrespective of the OTA and non-OTA solution, depending upon where the network-part of two-sided models are trained. Therefore, that should not be the focus of RAN2 discussion and argued by Apple and ZTE, RAN2 should focus on below.  **Alternative 1 (non-OTA approach):**  **NW dataset/model parameters collection entity** -> **UE training entity**  **Alternative 2 (OTA approach):**  **gNB** -> **UE** |
| Lenovo | No with comments | The first hop “gNB transfer dataset/model parameters” implies gNB itself is responsible of the decoder and virtual encoder training, which may not be the case.  For simple and less controversial discussion, **we suggest removing the first hop “gNB transfer dataset/model parameters” for both OTA and non-OTA based approach**. Then,   * **OTA:** NW-side dataset/model parameters collection entity (gNB/CN/OAM/gNB server) -> UE side training entity * **Non-OTA:** NW-side dataset/model parameters collection entity (gNB/CN/OAM/gNB server) -> UE   Besides, in our understanding for sure there will be intermediate node in the case of:   * gNB -> intermediate node -> UE/UE-side server * gNB server -> intermediate node -> UE/UE-side server   **Rapporteur response:**  Agree that gNB may itself be responsible for training, the alternatives above doesn’t preclude such case, i.e. NW-side dataset/model parameter collection entity could be (gNB/CN/OAM/gNB server) as shown in the figure |
| Mediatek | Yes for Alt.1 with comments;  No for Alt.2. | First of all, it’s better to align the terminology for 'OTT server', which has been clarified for UE-side data collection:   * OTT means the server is outside the MNO’s network, or * A server for UE-side model training, which is within the MNO’s network.   For Alternative 1, it can be revised as below:  **gNB** -> **NW dataset/model parameters collection entity** -> **UE training entity** (~~OTT~~ a server inside~~/outside of~~ MNO or an OTT server)  For Alternative 2, we are questioning why it is necessary to mention the NW dataset/model parameters collection entity for the OTA approach. Why are the directions between gNB and the NW dataset/model parameters collection entity different for non-OTA and OTA approaches?  Under the OTA approach, if the UE transfers the model parameters/dataset to the OTT server, similar to data collection option 1a, it is purely a UE implementation and out of 3GPP scope. Therefore, the only thing that needs to be discussed is the model parameter and/or dataset sharing from gNB to the UE.  If the UE transfers the model parameters/dataset to the server for UE-side model training within the MNO, why do these parameters/datasets need to be relayed by the UE over the air interface instead of the gNB directly sharing the model parameters/datasets with the server for UE-side model training within the MNO, assuming the server is inside the MNO?  Therefore, our understanding of Alternative 2 is:  gNB -> UE -> OTT server.  Since model parameter/dataset sharing from the UE to the OTT server is outside the scope of 3GPP, RAN2 only needs to discuss the feasibility of model parameter/dataset sharing from the gNB to the UE in Alternative 2. |
| Samsung | Please see comment | gNB is not the source of dataset/model parameters (but is rather the origin of the raw data set via data collection e.g. CSI measurements from UEs), unless gNB is training entity. Related to this issue, the figures state “CSI compression data collection a gNB”. In order to clarify this data is not compressed at the gNB, it would be better to remove the underlined word. The text related to the arrow linked to the transfer from gNB should say “raw CSI transfer”, and the text next to the NW-side entity should say “NW-side ~~dataset/model parameters~~ CSI collection entity”. |
| OPPO | Yes for Alt.1 with modification  No for Alt.2 | We understand rapporteurs’ intention is to high-level show how dataset and/or parameter sharing is workable from E2E solution point of view for both non-OTA and OTA direction. But the point is that which node hosting model training will impact the dataset and/or parameter sharing path as mentioned by Qualcomm and Lenovo. At this early stage, we should try to identity all the candidates on the table; otherwise, we may not have the whole picture for each direction, in this sense, we’d like to modify Alt.1 and Alt.2 like the following:  For non-OTA approach:  If dataset and/or parameter source is NG-RAN, the dataset and/or parameter sharing path is **gNB** -> **NW dataset/model parameters collection entity** -> **UE training entity;**  If dataset and/or parameter source is NW dataset/model parameters collection entity, the dataset and/or parameter sharing path is **NW dataset/model parameters collection entity** -> **UE training entity.**  For OTA approach:  If dataset and/or parameter source is NG-RAN, the dataset and/or parameter sharing path is **gNB** -> **UE** -> **UE training entity;**  If dataset and/or parameter source is NW dataset/model parameters collection entity, the dataset and/or parameter sharing path is **NW dataset/model parameters collection entity** -> **gNB** -> **UE** ->**UE training entity.** |
| Ericsson | Yes for Alt.1 (replace “gNB” with “NW-side data collection entity”, if needed)  Yes for Alt.2 (with modifications if needed) | Alt1: We do not agree with the proposal to remove “gNB” from the path. That is because 3GPP should also study how the dataset/model parameters reach the NW dataset/model parameters collection entity, not only how the dataset/model parameter are shared between the NW dataset/model parameters collection entity and the UE-server. For example, if the dataset/model parameters are generated by the NG-RAN, one option could be that the NG-RAN share this dataset/model parameters with the OAM (which in this case will be play the role of “NW dataset/model parameters collection entity”). If the “gNB” in the picture is not acceptable, we could replace the “gNB” in the picture with “NW-side data collection entity”. According to RAN2 discussion the NW-side data collection entity could be either the gNB or the OAM, hence it is natural to assume that the dataset/model parameters are generated by those entities. So that it will be: **NW-side data collection entity (gNB/OAM)**-> **NW dataset/model parameters collection entity** -> **UE training entity**  Alt2: The intention of this picture was to include all the possible options in which different NW entities could be involved, i.e 1) direct transfer from the gNB to the UE (in which case the gNB is both the NW-side dataset/model parameters collection entity and the entity in charge of the transferring), or 2) transfer from the NW-side dataset/model parameters collection entity to the UE (with indirect/direct involvement of the gNB). However, if 2) is not acceptable, 3GPP should at least discuss how the dataset/model parameters are shared between the NW-side data collection entity and the gNB (which in this case plays the role of the “dataset/model parameters collection entity”). For example, if the OAM does the NW-side data collection, and the OAM generates the dataset/model parameters, then the dataset/model parameters should be shared with the gNB before the gNB transmits them via OTA. So, we believe that just focusing on the path gNB->UE is not sufficient, given the options considered for the NW-side data collection. So we suggest the following rephasing, if needed.  **NW-side data collection entity (gNB/OAM)** -> **gNB** -> **UE** -> **UE training entity** |
| Nokia | No for Alt.1  No for Alt.2 | We agree with Apple that the discussion is limited to option 4 and option 3a-1.  **Alt 1 –** Because we haven’t been asked about how the NW acquires the datasets and parameter sets from the gNB, we do not think it is useful to include the gNB in the diagram for Alt 1, as suggested in Apple’s and Qualcomm’s comments.  **Alt 2** – Similarly, since we haven’t been asked to discuss and since it is out of our scope how the gNB acquires the datasets and parameter sets, the NW can be removed from the diagram, as suggested in Apple’s and Qualcomm’s comments.  While we do not agree with OTA dataset and parameter set transfer, we do agree with ZTE’s correction above.  We also agree with Samsung that the gNB isn’t likely to be the source of the datasets or parameter sets, but rather the source of raw data. However, we do not need to include that procedure of passing raw data from the gNB to the NW in our RAN2 discussions.  One aspect which has been exposed by Samsung’s comment is that we haven’t discussed whether the OTA solution uses CP or UP messages for the transfer. If the transfer is OTA, but between the UE and an entity connected to the AMF or to an application server in the NW, then the discussion is out of our scope. If the transfer is OTA between the UE and the gNB over RRC, then the discussion is in our scope, but we should ask ourselves why it would be desirable to consume crucial CP resources on dataset and parameter set transfer. |
| LGE | Alt1, Alt2 : No (Need modification) | We believe it is necessary to highlight the point that RAN2 should focus on, as mentioned by Apple and QC. This is because RAN1 LS only inquired about Option 3a-1 and Option 4, and RAN2 should only consider the scope of the issues raised. Therefore, the RAN2 study on OTA and non-OTA approaches should solely focus on the issues raised in the RAN1 LS.  **Alternative 1 (non-OTA approach):**  **NW dataset/model parameters collection entity -> UE training entity (OTT server inside/outside of MNO)**  **Alternative 2 (OTA approach):**  **gNB** -> **UE** |
| CMCC | No for Alt.1  No for Alt.2 | We agree with Apple that the discussion is limited to option 4 and option 3a-1.  For OTA and non-OTA approach, we are fine with Apple and QC’s suggestion:  **Alternative 1 (non-OTA approach):**  **NW dataset/model parameters collection entity** -> **UE training entity**  **Alternative 2 (OTA approach):**  **gNB** -> **UE** |
| Futurewei | No for Alt.1  No for Alt.2 | In RAN1 discussion, the two entities mentioned in the discussion are NW and UE, in which NW could include gNB, CN and OAM. We therefore agree with Apple and Qualcomm’s proposal. |

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. |

##### Summary:

First of all, based on companies’ feedback, though rapporteurs believe it would be helpful for companies to understand the overall procedure for dataset/model parameter transfer from data generation till received by UE-side OTT server, and then download trained model/information from OTT server to UE, rapporteurs also understand that RAN1 LS only focuses on dataset/model parameter transfer to UE-side OTT server. Therefore, in this email discussion, let’s only focus on dataset/model parameter transfer. However, Rapporteurs will leave the overall procedure and whether/how to concatenate dataset/model parameter path with model transfer path to TR rapporteur.

Regarding to the comment to remove the path between gNB to NW-side dataset/model parameter collection entity, as discussed in Q2-1, this is up to SA2/SA5 (RAN3). However, having such potential path in mind might also help us to identify any complexity and/or overhead when comparing solutions (e.g. Q2-9). For completeness and identifying potential impact to other WGs, rapporteurs suggest to capture this link in two alternatives for now, while RAN2 will only conclude on the feasibility analysis for the link where Uu interface is impacted, and leave other feasibility analysis to impacted WGs. Also, rapporteurs understand that the dataset/model parameter transfer between gNB and NW-side dataset/model parameters collection entity may not be needed based on different solutions, therefore, rapporteur further update such link as dot line.

Regarding to the comment to remove UE-UE side OTT server in Alternative 2 – first of all, as summarized above, Direction A asked by RAN1 LS requires offline engineering at UE OTT server. Therefore, based on rapporteurs’ understanding, the dataset/model parameters need to be sent to UE OTT Server in the end, no matter via OTA or non-OTA method. That’s why Alternative 2 also considers UE OTT server as the termination point. Furthermore, as discussed in Q2-11, this link could be similar to UE-side data collection. Considering there’s no additional effort from RAN2 at this stage (i.e. we will rely on conclusion to UE-side OTT server as discussed in Q2-11), rapporteur thinks it is fine to keep it for now for the completeness and align with RAN1 Direction A.

With above, two alternatives are updated as below:

Proposal 2: RAN2 to capture below two alternatives in TR for dataset/model parameter transfer Direction A:

|  |  |  |
| --- | --- | --- |
| **Alternative 1 (non-OTA approach):**  **gNB** -> **NW dataset/model parameters collection entity** -> **UE training entity** (a server inside/outside of MNO or an OTT server)   |  | | --- | | base station, cell tower, communication, connection, network, tower cloud, database, hosting, server cloud, server, web  **RAN2 analyzed area**  dataset/model parameters transfer  dataset/model parameter transfer  CSI compression data collection at gNB  NW-side dataset/model parameters collection entity (gNB/CN/OAM) 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** (if needed) **-> gNB -> UE** -> **UE training entity** (OTT server inside/outside of MNO)   |  | | --- | | iphone 14, iphone, mobile, smartphone, device, app, pro base station, cell tower, communication, connection, network, tower base station, cell tower, communication, connection, network, tower cloud, database, hosting, server cloud, server, web  **RAN2 analyzed area**  gNB  dataset/model parameter transfer  dataset/model parameter for training  CSI compression data collection at gNB  NW-side dataset/model parameters collection entity (gNB/CN/OAM) for two-sided model training  dataset/model parameter transfer  dataset/model parameters transfer  UE-side training entity for two-sided UE part model training |   **NOTE: The transfer path between gNB and NW dataset/model parameters collection entity (OAM/CN) in Alternative 1/2, if needed, is up to RAN3/SA2/SA5.** |

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. |
| Qualcomm | Agree (with modification) | We believe that for both solutions data transfer is needs between gNB to NW dataset/model parameters collection entity (OAM/CN/gNB server) considering   * In OTA, the data may need to be sent from the NW dataset/model parameters collection entity to gNB [when the network part of the two-sided model is trained at CN/OAM], and * In non-OTA, the data may need to be sent from gNB to the NW dataset/model parameters collection entity (when the network part of the two-sided model is trained at gNB).   Therefore, we suggest updating as below:  Standardization of the transfer procedures ~~from~~ between gNB to NW dataset/model parameters collection entity (OAM/CN/gNB server), if needed, is up to SA2/SA5.  We agree with Apple that RAN2 should not focus on the discussion of transfer procedures ~~from~~ between gNB to NW dataset/model parameters collection entity (OAM/CN/gNB server). |
| Lenovo | Yes | It’s ok to let SA2/SA5 to lead the relevant discussion. RAN3’s impact depends on where the dataset/parameter collection entity is and how is it connected to the gNB.  P.S. again, what transferred from gNB to NW dataset/model parameters collection entity may not be “dataset/model parameter” itself. It can be training data (e.g., CSI report and target CSI) for training the two sided model. |
| Mediatek | Partially Yes | For Alternative 1, the transfer path from the gNB to the NW dataset/model parameters collection entity, and finally to the server for UE-side model training, falls within the scope of SA2 and SA5.  For Alternative 2, we question the necessity of considering the NW dataset/model parameters collection entity. |
| Samsung | Yes but please see comment | We would like to clarify that gNB to NW server path is about raw data transfer and it is not the same as data set/model parameter that RAN1 asked us about. Nevertheless, it is correct that if raw data is transferred from gNB to NW server, it should be realm of SA2/SA5. |
| OPPO | Agree (with modification) | We share the similar view as Qualcomm, the node hosting model training will decide whether the dataset and/or parameter sharing is from gNB -> NW dataset/model parameters collection entity or vice versa. We also understand RAN3 may be involved if we’re talking about the interaction between gNB and CN/OAM. More addition, we never differentiate gNB server and gNB in the past, the necessity/motivation to differentiate gNB server and gNB is unclear, so we suggest the following:  **The discussion on dataset and/or parameter sharing procedure between gNB and NW dataset/model parameters collection entity (OAM/CN), if needed, may consult RAN3/SA2/SA5.** |
| Ericsson | Agree (OK with Oppo and QC modifications) | We agree with Oppo and QC proposal indicating that the standard procedures for the transferring of dataset/model parameters **between gNB and NW dataset/model parameter are up to RAN3/SA2/SA5.**  We also agree with Oppo that we do not need to mention the “gNB server”, since that is not an existing node from 3GPP point of view. We can adopt the same approach we used for NW-side model training when it was assumed that from 3GPP pov it does not matter if the training happens in the gNB or in a gNB-server. |
| Nokia | Yes |  |
| LGE | Yes with RAN3 | Since the interface between gNB and CN/OAM is involved, RAN3 needs be involved as well. Agree with the OPPO’s suggestion. |
| CMCC | Yes with comments | If CN is involved as NW dataset/model parameters collection entity, RAN3 is also involved for interface between gNB and CN. |
| Futurewei | Yes | Including RAN3, if CN is involved. |

##### Summary:

All companies (some agree with modification to add RAN3) agree that the transfer path from gNB to NW dataset/model parameters collection entity (OAM/CN/gNB server), if needed, is up to SA2/SA5. The proposal is merged in Q2-0.

### 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. |
| Qualcomm | 1. Maybe 2. Yes 3. Discussion not required. 4. Yes (with modification) | We do not need to separately consider the UE-side OTT server and UE server inside MNO. We can reuse the term used earlier, “**UE training entity (within/outside MNO).”**  We also agree with Apple that we should send LS to SA2/SA5 to check any issue.  For 1:   * Even when the training happens at the OAM, the dataset/parameters can be shared by the CN.   For 2:   * Based on the update by rapporteurs, the table needs to be updated  |  |  |  | | --- | --- | --- | | CN -> UE-side OTT server | ~~RAN3,~~ SA2 | ~~Up to RAN3 on NG impact,~~ SA2 |   For 3:   * We can leave it out of the discussion, as this can be left up to implementation.   For 4:   * If the training is happening at the gNB, then the parameters / dataset can be sent to UE side training entity via OAM/CN.   Therefore, we argue to update the table as below:     |  |  |  | | --- | --- | --- | | **Option** | **Impacted WG** | **Specification impact/Implementation impact** | | 1. OAM -> **UE training entity (within/outside MNO)** | SA5 | Up to SA5 | | 1. CN -> **UE training entity (within/outside MNO** | SA2 | SA2 | | 1. gNB -> **UE training entity (outside MNO)** | Outside of 3GPP | Up to implementation | | 1. gNB -> CN/OAM -> **UE training entity (within/outside MNO)** | RAN3, SA2, SA5 | RAN3, SA2, SA5 | |
| Lenovo | 1. Yes 2. Yes 3. No 4. No | 3) is strange. Does 3) eventually mean from “gNB server” to UE side OTT server?  We can leave 4) for the time being and focus on 2). |
| Mediatek | 1), 2) Yes | First of all, it’s better to align the terminology for 'OTT server', which has been clarified for UE-side data collection:   * OTT means the server is outside the MNO’s network, or * A server for UE-side model training, which is within the MNO’s network.   Therefore, 1), 2) can be revised as:   1. OAM -> The server for UE-side model training or OTT server 2. CN-> The server for UE-side model training or OTT server   Option 4) may be possible in implementation. But we don’t think we want to introduce a new standardized interface between the gNB and the server for UE-side model training. |
| Samsung | Please see comments | It is not necessary for RAN2 to identify various sub-options for this case, or the relevant WGs. RAN2 may not be aware of all the possible sub-paths within the Core anyway. The important thing to state is that RAN2 believe the non-OTA approach is not within RAN2 remit (from the Table above, it appears rapporteurs also feel this way), and so a clarification question needs to be asked to RAN1 on what aspect of non-OTA they felt was within RAN2 remit / they wanted RAN2 to comment on. |
| OPPO | 1. Yes with modification 2. Yes with modification 3. No   No | The similar view with ZTE, Option 3) and Option 4) is not practical without MNO involvement, so should be removed. gNB may be involved for dataset and/or parameter sharing as analysed in Q2-0, it should be one of the options, based on above, we suggest:     |  |  |  | | --- | --- | --- | | **Option** | **Impacted WG** | **Specification impact/Implementation impact** | | 1. OAM -> **UE training entity (within/outside MNO)** | SA5 | Up to SA5 | | 1. CN -> **UE training entity (within/outside MNO** | SA2 | Up to SA2 | | 1. gNB -> CN/OAM -> **UE training entity (within/outside MNO)** | RAN3, SA2, SA5 | Up to RAN3, SA2, SA5 | |
| Ericsson | 1),2),3) Yes  4) OK to replace it with: **OAM -> CN -> UE-side training entity** | 1),2),3) are straightforward and it can be considered. 3) is out of 3GPP, but it can be captured in the analysis (as we did for option 1a for UE-side model training).  The intention of 4) was to capture a possible alternative in which the dataset/model parameters are captured by a collection server within the MNO premises. We are ok to remove it with the option proposed by QC in Q2-3 below in which both the OAM/CN are involved. That is a feasible option from architectural point of view. |
| Nokia | 1) Yes  2) Yes  3) No  4) Maybe | A direct connection between a gNB and an OTT server outside the NW is out of 3GPP scope, so we cannot agree to discuss 3). Option 4 could be OK, but only until the point of the server inside the MNO. We agree that there are security and privacy concerns for 3) and 4).  We agree with Samsung that it is not necessary to provide these mappings. |
| LGE | 1) Yes  2) Yes  3) Yes  4) No | For 4) We share the same view with Samsung. We find it somewhat unclear why RAN2 is dividing the WG impact on non-OTA in such detail. RAN2 may not be fully aware of all the possible sub-paths within the Core anyway, and a clarification question needs to be posed to RAN1 regarding which aspects of non-OTA they believed fell within RAN2's remit and what they wanted RAN2 to comment on.  Additionally, for 4), the term "server inside MNO" is indeed ambiguous. It is important to clarify whether it refers to the UE-side OTT server within the MNO, the OAM server within the MNO, or both. The interpretation of "server inside MNO" will affect which working groups are impacted. |
| CMCC | 1. Yes 2. Yes 3. No 4. No | For option 3: We agree with ZTE that they are not allowed to directly transfer data from gNB to outside due to the security and privacy issue.  For option 4: If the server inside MNO resides at CN, we don’t see the difference between option 4 and option2. |
| Futurewei | 1. Yes 2. Yes 3. No 4. No | For option 3: agree with Nokia and other companies that a direction connection between a gNB and an OTT server outside the network is out-of-scope of 3GPP standard so we don’t need to discuss it.  For option 4: up to server inside MNO may be fine. The optional OTT server outside of MNO can be removed. |

##### Summary (covering Q2-3):

Regarding to SS’s comment on the necessarily of RAN2 analysis on above candidate solutions, it is rapporteurs’ understanding that we did the same analysis for UE-side data collection candidate solution, e.g. identifying candidate solutions from system level and identifying the impacted WGs. Also, since there’s no SI/TU for SA to study this aspect and RAN2 is the leading group for model transfer/delivery in general, similar as UE-side data collection, from RAN2 study point of view, we can propose several candidate solutions and further check feasibility to other WGs, then trigger the corresponding WGs for further study, similar as we did in UE-side data collection.

|  |  |
| --- | --- |
| **Options** | **Comments Summary** |
| 1. OAM -> UE-side OTT server | **Yes**: ZTE, Apple, HW, Vivo, QC, Lenovo, MTK, Ericsson, Nokia, LG, CMCC, FW  **Use ‘UE-training entity (inside/outside MNO)’:** QC, MTK, Oppo  [Rapp: since RAN2 use the term ‘UE-side OTT server’ for UE-side data collection, as proposed by MTK, rapporteurs further update it as UE-side OTT server or UE training entity (inside/outside MNO)] |
| 1. CN -> UE-side OTT server | **Yes**: ZTE, Apple, HW, vivo, QC, Lenovo, MTK, Ericsson, Nokia, LG, CMCC, FW  **Remove RAN3:** vivo, QC  [Rapp: agrees to remove RAN3, as we captured in Q2-1] |
| 1. gNB -> UE-side OTT server (outside of MNO) | **Yes**: HW, QC, Ericsson (ok to capture in analysis), LG  **Add RAN3:** vivo  **Security and privacy concern:** ZTE, Oppo (not practical without MNO involvement), Nokia, CMCC  **Implementation and no need to capture**: QC, Nokia, FW  **Check with SA5**: HW  [Rapp: Considering this is an implementation solution and some network vendors and operators have concerns on security and privacy, rapporteurs propose to remove it. However, as proposed by some companies, gNB may transfer dataset/model parameters via OAM/CN if gNB is responsible for NW dataset/model parameter collection entity. With that, rapporteurs further add ‘CN/OAM’ in between.] |
| 1. gNB -> server inside MNO -> optionally OTT server (outside of MNO) | **Yes**: QC  **Add RAN3:** vivo  **Security and privacy concern:** ZTE, Oppo (not practical without MNO involvement), Nokia, CMCC  **Ambiguous definition of server inside MNO**: Apple, LG  **No standardized new interface for gNB and server for UE-side training**: MTK  **Dataset/model parameter to be sent via OAM/CN:** QC, Ericsson  [Rapp: this solution is summarized from one contribution from RAN2 129 meeting, since the proponent is fine to consider other approach instead, rapporteurs here propose to remove this option.] |
| 1. OAM -> CN -> UE-side OTT server | **Yes**: QC, Ericsson  [Rapp: as commented by many companies, it might be possible that there’s different intermediate node between OAM and UE-side OTT server, which is up to SA5/SA2. Therefore, rapporteurs propose not to over-discussed non-OTA approaches and leave the intermediate node discussion to SA2/5. This also applies the same to Option 2] |

The updated candidate solution for ‘NW dataset/model parameters collection entity -> UE side OTT server (inside/outside MNO)’ can be found as below.

From RAN2 point of view, RAN2 identifies following candidate solutions for non-OTA solution ‘NW dataset/model parameters collection entity -> UE side OTT server/UE training entity (inside/outside MNO)’ , but no RAN2 impact is expected. We will further confirm feasibility in phase 2.

|  |  |  |
| --- | --- | --- |
| **Option** | **Impacted WG** | **Specification impact/Implementation impact** |
| 1. OAM -> UE-side OTT server or UE-side training entity (inside/outside MNO) | SA5 | Up to SA5  (any intermediate node between OAM and UE-side OTT server is up to SA5; CN involvement if needed is up to SA2/SA5 discussion) |
| 1. CN -> UE-side OTT server or UE-side training entity (inside/outside MNO) | SA2 | Up to SA2  (any intermediate node between CN and UE-side OTT server is up to SA2) |
| 1. gNB -> OAM/CN -> UE-side OTT server or UE-side training entity (inside/outside MNO) | RAN3, SA2, SA5 | Up to RAN3, SA2, SA5  (any intermediate node between gNB/OAM, OAM/UE-side OTT server, CN/UE-side OTT server is up to RAN3/SA2/SA5) |

##### 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** |
| Qualcomm | OAM -> CN -> UE-side training entity  [Even when the training happens at the OAM, the dataset/parameters can be shared by the CN. SA2 has procedures where the OAM exposes the data to NWDAF] | SA2 / SA5 | SA2 / SA5 |
| Ericsson | We are ok to capture the QC alternative above, since that it is an option that is viable from SA pov. |  |  |

##### Summary:

Please see the corresponding summary in Q2-2.

### 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 dataset/model parameters to UE via RRC signalling.

2) Solution 1b: gNB can transfer/deliver dataset/model parameters 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. |
| Qualcomm | Solution 1a Yes  Solution 1b  No | Same view as Apple and ZTE. |
| Lenovo | 1a: No  1b: No | Both are not really feasible in our view.  For 1a, RRC signalling is simply not scalable for dataset/parameter of size up to 255MB.  For 1b, to make it work, a new protocol layer handling the dataset/parameter segmentation/assembling would be probably required, which will have significant impact on existing protocol stack. |
| Mediatek | Yes | Both solution 1a and solution 1b can be evaluated for the feasibility study. |
| Samsung | Yes | We understand that solution 1b may require more specification work. But we don’t need to preclude it now, before we have further information from RAN1 in terms of data characteristics and gNB involvement. |
| OPPO | Solution 1a Yes  Solution 1b  Yes | Anyway, this is for analysis, not for normative work, if solution 1a/1b is considered for model transfer, we see no reason to exclude solution for study phase as the situation is quite similar between model transfer and dataset and/or parameter sharing. |
| Ericsson | 1a Yes  1b No | Both can be considered candidate solutions, and they can be captured in the analysis. However, solution 1b should not be further considered for the reasons mentioned by other companies above. |
| Nokia | 1a: No  1b: No | 1a: As was the case with model transfer, RRC signaling cannot support high data volume and it should not be extended to do so when there is a viable non-OTA method available.  1b: The complexity of terminating UP data in the gNB is too high considering that there is a viable non-OTA method available. |
| LGE | Yes | For the study phase, neither of the options do not need to be precluded. Feasibility of both options can be studied for supporting the transfer/delivery of super-size data set /model parameter. |
| CMCC | 1a: Yes  1b: No | Agree with Apple and ZTE. |
| Futurewei | 1) Solution 1a: Yes  2) Solution 1b: No | For Solution 1b, agree with ZTE, Apple and other companies that the mechanism (UP tunnel between gNB and UE) does not exist in current standard hence it is not feasible. |

##### Summary (covering Q2-7):

|  |  |  |
| --- | --- | --- |
| **Solution** | **Views on whether solution 1a as a candidate solution** | **Enhancements to be considered for feasible solutions/Reason of not feasible** |
| 1a | **Yes (12/14)**: ZTE, Apple, HW, vivo, QC, Oppo, MTK, SS, Ericsson, LG, CMCC, FW | * **Extend existing supported RRC signaling size (same spec impact of solution 1a in TR38.843)**: ZTE, Apple, HW, Vivo, QC, Oppo, MTK, SS, LG, CMCC |
| **No (2/14):** Lenovo, Nokia | * **not scalable for size up to hundreds of MB:** Lenovo, Nokia (shouldn’t extend when there’s a viable non-OTA method available) |
| 1b | **Yes (5/14)**: HW, Oppo, MTK, SS, LG (further study feasibility) | * **same as model transfer delivery 1b in TR38.843**: HW, Oppo, MTK, SS (understand more about data characteristics), LG |
| **No (9/14):** ZTE, Apple, vivo, QC, Lenovo, Ericsson, Nokia, CMCC, FW | * **no UP tunnel terminated between UE and gNB, where too much spec work is needed to a new protocol**: ZTE, Apple, vivo, QC, Lenovo, Ericsson, Nokia, CMCC, FW |

Regarding to solution 1a, 12/14 companies think it can be considered as a candidate solution at least from RAN2 point of view. Based on companies’ feedback during phase 1, please see detailed feasibility analysis in phase 2.

Observation: From RAN2 point of view, when gNB is the NW dataset/model parameter entity, OTA solution 1a (i.e. gNB -> UE via CP) can be considered as a candidate solution for dataset/model parameter transfer.

Regarding to Solution 1b, there’s a majority (9/14) think it is not feasible from RAN2 point view, considering the tremendous specification impact. Recalling that, during Rel-18, we have analysis following gaps/specification impact in TR38.843:

|  |  |  |
| --- | --- | --- |
| **Current status and Gaps of Solution 1b in TR38.843** | **Potential RAN specification impact of Solution 1b in TR38.843** | **Applicable to dataset/model parameter transfer?** |
| - No model size limitation  - PDU session termination at gNB is not supported | Requires PDU session termination at gNB if needed | **Yes (A1-Size)** |
| Model transfer continuity if PDU session terminated at gNB is not studied | - Identify a solution to support service continuity support between gNBs when PDU session is terminated at gNB if needed  - Xn/NGAP enhancement(s) for model transfer/delivery continuity | **Yes with change (A2-Continuity):**  - Identify a solution to support service continuity for dataset/model parameter transfer during UE mobility support between gNBs when PDU session is terminated at gNB if needed  - Xn/NGAP enhancement(s) for dataset/model parameter continuity |
| Management and interaction between UE and gNB appear to be feasible but not supported | Requires management and interaction between UE and gNB (e.g., model identification, model transfer completion indication, etc.) when model management at gNB | Not applicable since it’s not model transfer |
| - Procedure latency depends on model size, QoS requirement and DRB priority  - QoS management at gNB if PDU session is terminated at gNB is not supported | Identify a solution to support QoS management at gNB for model transfer when PDU session is terminated at gNB if needed | **Yes (A4-Latency)** with change to dataset/model parameter transfer |

It is observed that whether PDU session can be terminated at gNB needs to be further checked with SA2, and also this solution might require new protocol stack. Considering the significant specification impact to support establishing UP tunnel between gNB and UE, a huge effort is needed across different WGs to support solution 1b in 5GA, which is not practical.

Based on above analysis,

Proposal 3: From RAN2 point of view, when gNB is the NW dataset/model parameter entity, solution 1b (i.e. gNB -> UE via UP) is not considered as a candidate solution in 5GA.

##### 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** |
| Mediatek | gNB->UE | A specific radio bearer dedicated to model parameter and data transfer, terminated between the UE and NG-RAN, combines the advantages of SRB (data content visibility to NG-RAN) and DRB (unrestricted data amount). | RAN2, SA2, RAN3 |
|  |  |  |  |

##### Summary:

In rapporteurs’ understanding, the proposed solution is solution 1b, where a UP tunnel between UE and gNB is needed.

#### 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. |
| Qualcomm | 1. Yes 2. No 3. No | If the dataset/parameters are available at the OAM/CN, then they can be shared using a non-OTA approach. No need to unnecessarily complicate the procedure. |
| Lenovo | 2) Yes  3)Yes | 1) Can be probably excluded as commented in Q2-6 |
| Mediatek | 1) Yes  2) No  3) No | We share the same understanding as HW. For OTA approach, the only data path needs to be considered is gNB-> UE. |
| Samsung | Please see comment | The difference between ‘gNB 🡪 UE’ and ‘NW 🡪 UE’ appears to be that in the latter case (i.e. for cases 2) and 3) immediately above), the solution does not require any changes to the Uu link i.e. regardless of whether ‘NW 🡪 UE’ uses CP or UP, the transfer from gNB to UE is using existing mechanisms on the radio link. So we essentially agree with the rapporteur analysis.  We would like to reiterate that in our understanding, data set/model parameters are generated in NW training entity. So, the path should strictly speaking be NW training entity (NW server) 🡪 gNB 🡪 UE 🡪 UE OTT server. CN/OAM are involved in the NW server to gNB transfer part of the procedure, which is the same as non-OTA case.  For case 1), we foresee a case where a gNB1 may configure CSI measurement but gNB2 may train the model after receiving CSI data from gNB1. If companies agree that this is a valid case, RAN3 should also be involved. |
| OPPO | 1) Yes  2)Yes  3) Yes | Anyway, this is for analysis, not for normative work, we see no strong reason to exclude solution for study phase as the situation is quite similar between model transfer and dataset and/or parameter sharing. |
| Ericsson | 1) Yes  2) No  3) Yes (at least for the case in which the OAM is the NW-side data collection entity) | For 2) we are ok to remove it from the possible options. That is because we have not studied before how the NW-side collected data (either by the gNB or by the OAM) can become available at the CN. Hence, before considering this option we should discussed that. Considering the complexity we are ok to rule out this option.  For 3), we believe that this option should be considered, at least for the case in which the OAM is the NW-side data collection entity. In this case, it is natural to assume that the OAM generates the dataset/model parameters, and it passes them to the gNB for the OTA delivery. If it is needed, it can be clarified that in this case, the OAM is the NW-side data collection entity and the gNB is the NW dataset/model parameters collection entity. |
| Nokia | See Comment | The WG impacts are correct, but we do not want to imply agreement with the concept of using UEs as relays to transfer data from the NW to servers for training UE-side models. |
| LGE | 1) Yes | For 2) and 3), dataset/parameters available at OAM/CN can be shared by both non-OTA and OTA approach. We think the discussion on what procedure to send it through is more appropriate for another WG than RAN2. |
| CMCC | 1. Yes 2. No 3. No | For OTA approach, gNB-> UE is the only data path to be considered. |
| Futurewei | See comments | For all options, we think the answer is Yes for model parameters for inference purpose at the UE, and No for training dataset, assuming the dataset will then be transferred to model training server. |

##### Summary (covering Q2-10):

|  |  |  |
| --- | --- | --- |
| **Option** | **Views on whether can be considered as a candidate solution** | **Enhancements to be considered for feasible solutions/Reason of not feasible** |
| 1. gNB -> UE | **Yes (12/14)**: ZTE, Apple, HW, Vivo, MTK, Oppo, Ericsson, SS, LG, CMCC, FW (yes for model parameter) | * **same as Q2-6** |
| **No (2/14):** Lenovo, Nokia | * **same as Q2-6** |
| 1. CN -> gNB -> UE | **Yes (6/14):** ZTE, vivo, Lenovo, Oppo, SS, FW (yes for model parameter) | * **only consider when dataset/model parameter is generated NW training entity (CN)**: SS |
| **No (8/14):** Apple, HW, QC, MTK, Ericsson, Nokia, LG, CMCC | * **dataset/model parameter is transparent to gNB:** Apple   [Rapp] Update it as CN -> UE via gNB, same for solution 3.   * **no benefit compared to non-OTA approach as dataset/model parameter is generated by gNB**: HW, QC, MTK * **haven’t study feasibility of CN as NW training entity:** Ericsson * **leave it to other WGs:** LG |
| 1. OAM -> gNB -> UE | **Yes (7/14)**: ZTE, vivo, Lenovo, Oppo, SS, Ericsson | * **only consider when dataset/model parameter is generated NW training entity (OAM)**: SS, Ericsson |
| **No (6/14):** Apple, HW, QC, MTK, Nokia, LG, CMCC | * **dataset/model parameter is transparent to gNB:** Apple * **no benefit compared to non-OTA approach as dataset/model parameter is generated by gNB:** HW, QC, MTK * **leave it to other WGs:** LG |

Considering the diverged views on OTA solution 2 (i.e. CN -> UE via gNB) and OTA solution 3 (i.e. OAM -> UE via gNB), let’s discuss the feasibility/complexity of above two solutions in phase 2, then decide whether to consider above two solutions as candidate feasible solution. Detailed discussion can be found in phase 2.

##### 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** |
| Ericsson | OAM (NW-side data collection entity)-> gNB (NW dataset/model parameters collection entity)-> UE | If Option 3) in Q2-9 above is removed, we should at least consider the case in which the OAM is the NW-side data collection entity. In such case, the NW dataset/model parameters can be generated by the OAM, transferred to the gNB, and then transmitted by the gNB via OTA. |  |
|  |  |  |  |

##### Summary:

Please see the corresponding summary in Q2-9.

#### 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. |
| Qualcomm | Agree for option 1a) and 1b).  No for 2 and 3. | OTA solution should be discussed only in the context of options 1a) and 1b).  In solution 2, the data from the UE is sent to the CN; thereafter, the data is sent from the CN to the UE server (within/outside the MNO network). In solution 3, the data is sent to OAM; thereafter, the data is sent from CN to UE server (within/outside MNO network). Therefore, options 2 or 3 should not be candidates for the OTA approach for dataset and parameter sharing. |
| Lenovo | 1a yes | For 1b), 2, 3, if the dataset/parameter will be terminated in 3GPP NW as intermediate step, it is no difference (maybe only drawback) compared to non-OTA based approach. |
| Mediatek | Only option 1a | It should be clarified that the OTT server in question is an OTT server outside of the MNO, not a server within the MNO.  As commented in Q2-0, if the UE transfers the model parameters/dataset to a server for UE-side model training within the MNO, why do these parameters/datasets need to be relayed by the UE over the air interface? Instead, the network side could directly share the model parameters/datasets with the server for UE-side model training within the MNO, assuming the server is inside the MNO.  Therefore, for the OTA approach, only option 1a is reasonable, where the UE transfers the model parameters/dataset to the OTT server.  However, under the OTA approach, model parameter/dataset transfer from UE to the OTT server is purely a UE implementation similar to data collection option 1a, it is out of 3GPP scope. |
| Samsung | OK with 1a | Agree with Huawei in that – for Rel-19 – only option 1a is supported. |
| OPPO | 1a only | First of all, 1b is precluded by RP for UE-sided data collection, no need to consider this option in this email, as for 2) and 3), the similar view as vivo, it has logic problem to consider 2)/3) as 2)/3) means in DL CN/OAM transfer the data to UE and then UE transfer data back to CN/OAM in UL, why CN/OAM would like to do something like this? So 2)/3) should not be considered. |
| Ericsson | 1a,1b,2 (following the discussion on UE-side data collection) | If we assume that only the gNB or the OAM can generate the dataset/model parameters, then from the RAN2 point of view, the dataset/model parameters can be transmitted by the UE via any of the options studied for UE-side data collection (except option 3 since the OAM was already involved in the procedure to transfer the dataset/model parameters to the UE). This means that 1a, 1b, 2 could be in principle re-used. |
| Nokia | No | We do not agree with option 1b since it is over the air, and we cannot object to option 1a since it is out of 3GPP scope, and we will thus also not agree to support it. Options 2 and 3 have the UE sending the data it just received from the network back to the network to transfer to the UE-side server. That is inefficient compared to not involving the UE. |
| LGE | Yes for Option1a, comments for other options. | It is straightforward to agree Option1a for Rel-19.  For Option2 and Option3, LS to other WGs(SA2, SA5, etc) would be needed for validity check of whether option 2 and option 3 of UE-sided data collection can be reused to two-sided model. |
| CMCC | Only 1a | For OTA approach, it is straightforward that option 1a can work. And for other options, we share similar view as vivo. |
| Futurewei | Yes for 1a | For OTA approaches, Option 2 and 3 are not reasonable solutions. Option 1b has been removed from the list of viable options, at least for R19. |

##### Summary:

|  |  |  |
| --- | --- | --- |
| **Option** | **Views whether to consider as candidate solution** | **Enhancements to be considered for feasible solutions/Reason of not feasible** |
| 1a | **Yes (14/14)**: ZTE, Apple, HW, vivo, QC, Lenovo, MTK, SS, Ericsson, Oppo, Nokia, LG, CMCC |  |
| **No:** |  |
| 1b | **Feasible (2/14): QC, Ericsson** |  |
| **Not feasible/not preferred as candidate solution (12/14):** ZTE, Apple, HW, vivo, Lenovo, MTK, SS, Oppo, Nokia, LG, CMCC, FW | * **for Rel-19, only 1a is supported:** HW, SS |
| 2 | **Feasible (1/14)**: Ericsson | * **in principle reused considering OAM as NW dataset/model parameter entity:** Ericsson |
| **Not feasible/not preferred as candidate solution (13/14):** ZTE, Apple, vivo, QC, Lenovo, MTK, SS, Oppo, Nokia, LG, CMCC, FW | * **wait for SA2:** Apple * **same information over air interface again:** HW |
| 3 | **Feasible ():** |  |
| **Not feasible/not preferred as candidate solution (14/14):** ZTE, Apple, vivo, QC, Lenovo, MTK, SS, Oppo, Nokia, LG, CMCC, FW | * **wait for SA2:** Apple * **same information over air interface again:** HW |

There’s a clear majority prefers only consider option 1a for the link between UE to UE-side OTT server for dataset/model parameter transfer. Furthermore, considering SA2/5 is studying solution 2 and 3, rapporteurs think we can first conclude on solution 1a and wait until SA2/5 complete their study for solution 2/3.

Proposal 4: UE transfer the received dataset/model parameter to UE training entity (OTT server inside/outside of MNO) transparently to 3GPP network.

# Phase 2 Discussion

After phase 1 discussion, rapporteurs believe companies now have better understanding on how each solution works. Furthermore, rapporteurs also further provide feasibility and complexity analysis based on companies’ feedback during phase 1. During phase 2, let’s focus on whether phase 1 proposals are acceptable and further discuss the feasibility and complexity of candidate solutions after phase 1.

## Agreeable Proposals:

Principle (new principles proposed in phase 1 to be discussed in Q3-2)

Proposal 1: Dataset/model parameter transfer solution from NW to UE training entity shall follow below principles:

* **A1 -** **Size**: From RAN2 point of view, RAN2 aims to define a unified solution to support various sizes of dataset/model parameter transfer (dataset and/or parameter sharing size could range from tens of KBs to hundreds of MBs, but in average around hundreds of MBs);
* **A2 - Continuity**: Service continuity of dataset and/or parameter transfer/delivery during UE mobility needs to be supported;
* **A3 - Controllability**: NW controllability: Decision on if and when to transfer/delivery the dataset and/or model parameter from NW to UE training entity (OTT server inside/outside of MNO);
* **A4 - Latency**: Relaxed latency requirement and infrequent update;
* **A5 - Visibility**: standardized data format and model structure for dataset and parameter to be understandable by UE.

*Alternatives for capturing in TR38.843:*

Proposal 2: RAN2 to capture below two alternatives in TR for dataset/model parameter transfer Direction A:

|  |  |  |
| --- | --- | --- |
| **Alternative 1 (non-OTA approach):**  **gNB** -> **NW dataset/model parameters collection entity** -> **UE training entity** (a server inside/outside of MNO or an OTT server)   |  | | --- | | base station, cell tower, communication, connection, network, tower cloud, database, hosting, server cloud, server, web  **RAN2 analyzed area**  dataset/model parameters transfer  dataset/model parameter transfer  CSI compression data collection at gNB  NW-side dataset/model parameters collection entity (gNB/CN/OAM) 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** (if needed) **-> gNB -> UE** -> **UE training entity** (OTT server inside/outside of MNO)   |  | | --- | | iphone 14, iphone, mobile, smartphone, device, app, pro base station, cell tower, communication, connection, network, tower base station, cell tower, communication, connection, network, tower cloud, database, hosting, server cloud, server, web  **RAN2 analyzed area**  gNB  dataset/model parameter transfer  dataset/model parameter for training  CSI compression data collection at gNB  NW-side dataset/model parameters collection entity (gNB/CN/OAM) for two-sided model training  dataset/model parameter transfer  dataset/model parameters transfer  UE-side training entity for two-sided UE part model training |   **NOTE: The transfer path between gNB and NW dataset/model parameters collection entity (OAM/CN) in Alternative 1/2, if needed, is up to RAN3/SA2/SA5.** |

OTA solution

Proposal 3: From RAN2 point of view, when gNB is the NW dataset/model parameter entity, solution 1b (i.e. gNB -> UE via UP) is not considered as a candidate solution in 5GA.

UE -> UE training entity (OTT server inside/outside of MNO)

Proposal 4: UE transfers the received dataset/model parameter to UE training entity (OTT server inside/outside of MNO) transparently to 3GPP network.

#### Q3-1: Do companies agree with above proposals? (Companies who agree with proposals don’t need to provide feedback to this question)

|  |  |
| --- | --- |
| **Company (only if you don’t agree with above proposals)** | **Comment** |
| Apple | Agree P3. Comments on other 3 proposals.   * P1: On A5, we don’t think it is technically correct to change “**open** data format and model structure” in original Rapporteur proposal to “**standardized** data format and model structure”:   + “**Standardized** data format…”: we think it means 3GPP will introduce a new data/model format in RRC   + “**Open** data format ...” means data/model format of existing open source in industrial (e.g., ONNX, h5) can also be reused, besides 3GPP introducing a new data/model format.   As it is common understanding in RAN1/RAN2 that AI/ML algorithm is transparent to 3GPP, we think it seems not reasonable for 3GPP to introduce a new data/model format. At least, open data/model format should not be precluded at this stage. To make progress, we suggest to either add “open/” before standardized or remove “standardized”. As the key point of A5 is that UE can comprehend data/parameter, we slightly prefer later:  **A5 - Visibility**: **~~standardized~~** data format and model structure for dataset and parameter to be understandable by UE.   * P2:  1. Explanation of “UE training entity” for Alt-1 and Alt-2 are not same and there is overlapping (e.g. we think a server outside of MNO is exactly OTT server). To align wording, we suggest to use same wording “(a server inside of MNO or an OTT server)” as explanation of “UE training entity” for both Alt-1 and Alt-2:   **Alternative 1 (non-OTA approach):**  **gNB** -> **NW dataset/model parameters collection entity** -> **UE training entity** (a server inside**~~/outside~~** of MNO or an OTT server)  **Alternative 2 (OTA approach):**  **gNB** -> **NW dataset/model parameters collection entity** (if needed) **-> gNB -> UE** -> **UE training entity** (~~OTT server inside/outside of MNO,~~ a server inside of MNO or an OTT server)   1. In Figure of Alt-1, the path between gNB and data/parameter collection entity should also be dotted line, if we follow the same logic of figure in Alt-2 (i.e. because the data/parameter collection entity can be gNB).  * P4: First, the explanation of UE training entity has some issue (OTT server is outside of MNO according to its name Over the Top). We think it can be removed as we have captured it in P2. Secondly, this proposal is only applied to OTA approach (rather than non-OTA), which should be clarified.   Proposal 4: In OTA approach, UE transfers the received dataset/model parameter to UE training entity ~~(OTT server inside/outside of MNO)~~ transparently to 3GPP network. |
| vivo | **P1**: Comment:  We wonder how to understand “unified solution”. From R1 LS the typical size of model parameters is 11.6MB and dataset size is 225MB. From that point “*average around hundreds of MBs*” is not fully aligned with R1 Ls. How to consider “unified solution” in case the dataset is send to multiple UEs? From our understanding, in case of model parameters transfer there may not any necessity of sending the model parameters to more than one UEs.  **P2**: No for P2  These two alternatives are used to align the companies understanding on the potential solutions and impact WGs, and no need to capture the figures in the TR. E.g., the scope marked as ‘RAN2 analysed area’ is ambiguous if captured in the TR.  **P3**: “**…when gNB is the NW dataset/model parameter collection entity….**?  **P4**: Comment  Similar concern as Apple, OTT server is outside of MNO. |
| Lenovo | P1-A3: It reads now only applicable to non-OTA scenario, which is not the intention? suggest clarification, e.g.,  NW **~~controllability: Decision on~~ decides** if and when to transfer/delivery the dataset and/or model parameter from NW to **UE or** UE training entity (OTT server inside/outside of MNO);  P2: As commented in Q2-0,the first hop “gNB transfer dataset/model parameters” implies gNB itself is responsible of the decoder and virtual encoder training, which may or may NOT be the case.   * If gNB is responsible of the decoder and virtual encoder training, then gNB is considered as the dataset/model parameters collection entity as already covered in the figure “NW-side dataset/model parameters collection entity (**gNB**/CN/OAM/gNB server)” * If NOT, then what transferred from gNB to the NW-side dataset/model parameters collection entity is only raw training data (e.g., CSI before or after compression), which is different than “data set/model parameters”.   In either case, we can simply remove the first hop gNB. It’s enough to consider gNB as one possible NW-side dataset/model parameters collection entity as the second hop captured in the figure now. |
| Huawei, HiSilicon | **For the above two figures, we have some comments:**  (1) We think that RAN1 LS is focusing on dataset/parameter related to encoder, i.e. UE-part of a two-sided CSI compression model, and NW-part is out of RAN1 LS's scope. In this case, we understand that the collection entity is just a relay for UE-part dataset/parameter.  In current figures, it is using "NW-side dataset/model parameters collection entity (gNB/CN/OAM) for two-sided model training", and it may lead to some confusions like:   * this entity can train UE-part model * this entity can train NW-part model   In order to focus on the scope mentioned in the RAN1 LS, we suggest to simplify the naming by using:  **data collection entity for UE-part dataset/parameter**  (2) For the Note under figures, we do not think it will impact feasbility of OTA and non-OTA approaches, so we suggest to clarify it.  **NOTE: The transfer path between gNB and NW dataset/model parameters collection entity (OAM/CN) in Alternative 1/2, if needed, is up to RAN3/SA2/SA5, and this does not impact feasiblity of OTA and non-OTA approaches.**  **For principles, we have some comments:**  **For A1 - Size**, we can understand the intention, but we think here "a unifed solution" can be improved, it can be a mixed solution of OTA and non-OTA (e.g. NW can choose to select one of them based on data volume and other factors). In this case, we suggest to improve the wording a bit:   * **A1 -** **Size**: From RAN2 point of view, RAN2 aims to define a unified solution (e.g. OTA, non-OTA or a combination) to support various sizes of dataset/model parameter transfer (dataset and/or parameter sharing size could range from tens of KBs to hundreds of MBs, but in average around hundreds of MBs);   **For A2 - Continuity**, we understand that the dataset/parameter should be transferred from NW to UE-side completely, so the the object should be dataset/parameter rather than a specific UE. In this case, we suggest to improve the wording a bit:   * **A2 - Continuity**: Service continuity of dataset and/or parameter transfer/delivery during UE mobility needs to be supported (this is per dataset/parameter); |
| ZTE | P1: In P1, ZTE does not agree with the aiming of unified solution defined in A1, firstly, no matter the OTA solution or non-OTA solution,we do not preclude any one of them so far as now, at least we have two solutions on the table basically. Secondly, RAN2 only can study the OTA solution and have no ability to analyze and evaluate the non-OTA solution. In this sense, we do think RAN2 can make a conclusion to aim a unified solution, so that we propose the modify A1 like below:  Proposal 1: Dataset/model parameter transfer solution from NW to UE training entity shall follow below principles:   * **A1 -** **Size**: The solutions shall be to support various sizes of dataset/model parameter transfer (dataset and/or parameter sharing size could range from tens of KBs to hundreds of MBs, but in average around hundreds of MBs); * **A2 - Continuity**: Service continuity of dataset and/or parameter transfer/delivery during UE mobility needs to be supported; * **A3 - Controllability**: NW controllability: Decision on if and when to transfer/delivery the dataset and/or model parameter from NW to UE training entity (OTT server inside/outside of MNO); * **A4 - Latency**: Relaxed latency requirement and infrequent update; * **A5 - Visibility**: standardized data format and model structure for dataset and parameter to be understandable by UE.   P2: We agree to capture something in the TR, but there is no need for RAN2 to capture the whole picture for each solution as RAN1 only mention RAN2 is just to evaluate how to transfer the data from NW side to the UE side, how data transfer within the NW is another story. In this sense, RAN2 only is suggested to capture the RAN2 analysis area in the TR and leave the other possible areas to the other working group in the WI (e.g. RAN3 or SA group)  For OTA based solution:   |  | | --- | | iphone 14, iphone, mobile, smartphone, device, app, pro base station, cell tower, communication, connection, network, tower  dataset/model parameter transfer  gNB  UE |   For NON-OTA based solution:   |  | | --- | | cloud, server, web cloud, database, hosting, server  dataset/model parameter transfer  UE-side training entity for two-sided UE part model training  NW-side dataset/model parameters collection entity (gNB/CN/OAM) for two-sided model training | |  | |
| Qualcomm | Agree with Apple changes related to P2 and P4.  **Further comment on P2:**  The NOTE is not clear.  NOTE: The transfer path between gNB and NW dataset/model parameters collection entity (OAM/CN) in Alternative 1/2, if needed, is up to RAN3/SA2/SA5.  We prefer to update the NOTE as below:  **NOTE: The standardization of dataset/model parameters transfer between gNB and NW dataset/model parameters collection entity (OAM/CN) in Alternative 1/2, if needed, is up to RAN3/SA2/SA5.**  **Comment on P1:**  We agree with Apple, but it need further change, as there is no model structure for dataset.  **A5 - Visibility**: **~~standardized~~** data format and model structure for **~~dataset and~~** parameter to be understandable by UE |
| OPPO | P1： We understand P1 is just for requirements not for specific solution, the terminology ‘unified’ is more like a solution, so to make it generic, we suggest the following:   * **A1 -** **Size**: Supporting various sizes of dataset/model parameter transfer (dataset and/or parameter sharing size could range from tens of KBs to hundreds of MBs, but in average around hundreds of MBs);   P2/P4: We’re fine with Apple’s suggestion. |
| Ericsson | **P1:**   * On A1, as a principle 3GPP should aim at defining a unified solution, no matter if eventually that will be OTA or non-OTA based. This is what typically 3GPP tries to do, i.e. since both approaches have the same objective (which is how to transfer the dataset/model parameters to the training entity) we should avoid defining alternative solutions for the same objective to avoid unnecessary complications in the specifications (especially since many different WGs would need to be involved depending on whether the solution will be OTA or non-OTA). * A5: we are ok with the change proposed by Apple.   **P2:** We are ok with the change proposed by QC to the NOTE. We are also ok with the proposal by Apple to replace server outside the MNO with OTT server.  **P3:** OK  **P4:** We partly disagree with Apple proposal. For P2, it is proposed that the UE training entity can be the server inside the MNO or OTT server (outside the MNO). We need to stick to this definition in both OTA and non-OTA, we cannot change the definition based on the specific solution. Hence, we propose the following:  Proposal 4: In OTA approach, UE transfers the received dataset/model parameter to UE training entity ~~(OTT server inside/outside of MNO~~ a server inside of MNO or an OTT server~~)~~ transparently to 3GPP network. |
| Nokia | Our answer is in the context of RAN2, which is between the UE and the gNB.  **P1 – A1: OK, A3: OK, A4: OK – However, we are not agreeing that RAN2 should be involved in the dataset and parameter set transfers from the NW to server for training UE-side models.**  A2 – **Not OK.** We do not agree that service continuity should be supported through RAN2 means. The purpose of the dataset and parameter set transfers being discussed are to provide the entity training UE-side models with useful data to train UE-side models. As with other protocols which run atop the 3GPP network, we should push the service continuity aspects to other layers.  A5 – **Not OK.** We see no reason for the UE to be able to interpret the datasets and parameter sets being transferred to the entity which trains UE-side models. The UE is not training the UE-side models, so the transfer can be completely transparent to the UE. Additionally, it isn’t clear that the parameter set would be compatible with UEs since the parameter set represents an encoder model that can be run on a gNB and a vendor server, not on a UE modem.  **P2 – Alternative 1: OK, however out of our scope. We should respond to the LS with this information instead of putting it in the RAN1/RAN2 TR.**  **Alternative 2: Not OK. We do not think it is in our scope to discuss how the datasets or parameter sets arrive at the gNB and isn’t important to RAN2 discussion.**  We should focus on the interaction between the gNB and the UE. We can’t really comment on the UE to UE-side training entity component either, since it appears to be proprietary.  For completeness, and to demonstrate the potential for inefficiency, we would like to add that it could be the gNB between the UE and the UE-side Training Entity. It would be clear in that case that the air resources would be used twice for each transfer, once in the UL and once in the DL.  **P3 – OK. But we also think that the solutions 1a, 1b, etc., which we came up with are unrelated to this discussion. Those solutions are for transferring executable, trained models to the UE, not for using UEs as relays for data transfers between the NW and other non-UE entities.**  **P4 – Not OK. How can we ensure security when the transfer from the UE to the UE-side training entity doesn’t guarantee a secure transfer?** |
| CATT | P1-P3 look good to us.  P4: Agree with Apple to add “In OTA approach” as restrictive condition. |
| Mediatek | **P1**:  We should adhere to the terminology used in the RAN1 LS. Additionally, the model structure and model parameters are distinct aspects of describing an AI model. Further discussion is needed to determine how to achieve visibility of the model parameters and dataset. RAN1 assumes that the model parameters and dataset are understandable between the UE-side and the network-side. However, RAN1 does not assume that this visibility is solely achieved through standardization, as offline engineering is assumed during the two-sided model offline training process.  **A5 - Visibility**: ~~standardized data format and model structure for~~ dataset and model parameter to be understandable by UE.  **P2：**Agree with Apple.  **P3**: OK  **P4:**  Just as commented in Q2-0, Under the OTA approach, if the UE transfers the model parameters/dataset to the OTT server, similar to data collection option 1a, it is purely a UE implementation and transparently to 3GPP network.  However, if the UE transfers the model parameters/dataset to a UE training entity within the MNO, why do these parameters/datasets need to be relayed by the UE over the air interface instead of the gNB directly sharing the model parameters/datasets with the UE training entity within the MNO.  Even if considering UE transfers the received dataset/model parameter to UE training entity inside MNO, it should not be transparent to 3GPP network.  Therefore, P4 should be revised as:  P4: UE transfers the received dataset/model parameter to UE training entity (OTT server ~~inside/outside of MNO~~) transparently to 3GPP network. |
| Futurewei | **P2: No to Alt. 2 and its figure**. In this approach, dataset was collected at the NW-side and transferred to the UE via gNB. The UE then transfers it back to the gNB again (i.e., OTA) and the dataset eventually goes to the UE-side training entity. This is not the smart way of implementation and it increases network traffic. |
| LGE | For P1-A5, we think the clarification is necessary as follows :   * **A5 - Visibility**: standardized data format and**/or reference** model structure for dataset and parameter to be understandable by UE.   The directions/options below are rooted in the RAN1 agreement below.   |  | | --- | | **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.*** |   As you can see, Option 3 requires a standardized reference model structure and Option 4 requires a standardized data/dataset format. The reference model is what is being standardized here, which means that it does not have to exactly match the UE/NW side model.  For P2, it is necessary to align the wording between P2 and section 3.3. The lack of consistent terminology for the same entity can lead to confusion. It is important to ensure that both sections use the same expression to avoid misunderstandings.   |  |  | | --- | --- | | P2 | Section 3.3 | | **UE training entity** (a server inside/outside of MNO or an OTT server) | UE-side OTT server or UE-side training entity (inside/outside MNO) | | **UE training entity** (OTT server inside/outside of MNO) |   Therefore, we propose the following expressions to ensure consistency in terminology.  **Alternative 1 (non-OTA approach):**  **gNB** -> **NW dataset/model parameters collection entity** -> **UE-side OTT server or UE-side training entity (inside/outside MNO)**  **Alternative 2 (OTA approach):**  **gNB** -> **NW dataset/model parameters collection entity** (if needed) **-> gNB -> UE** -> **UE-side OTT server (inside/outside MNO)** |
| Samsung | P1: OK as general guidelines; practical usefulness of some of the principles is unclear, especially A4, but we can keep it  Also, A3 appears incomplete – the intended meaning is presumably that the decision is down to the NW?  P2: general comment – we appreciate that ‘RAN2 analyzed area’ is introduced by the rapporteurs to show the scope of our discussion, and highlight the fact that RAN2 does not need to discuss the transfer of raw data (which we agree with); however we maintain that we should say ‘CSI ~~compression~~ data’ (or at least ‘data related to CSI compression’), and also that – for the case where the NW entity is not at the gNB – we should not state that the gNB to NW-side entity link carries dataset/model parameters (whereas in fact dataset/model parameters are updated using the collected data).  Additionally, it is not the entire shaded area that will be ‘analysed’ by RAN2 in the sense that RAN2 will not develop an e2e solution (perhaps ‘analysed’ can be considered broad enough)  P3 – ok in principle but does this not violate the principle of a unified solution?  P4 – ok |

## Additional Principle

#### Q3-2: Do companies agree to additionally consider below principles on top of Proposal 1?

1. Uu interface impact (e.g. radio resource)
2. UE impact (e.g. power/ memory requirement)
3. UE selection for being reception of dataset/model parameter
4. E2E reliability of dataset/model parameter transfer
5. proprietary information of NW topology and NW vendor should be respected and not disclosed.

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comment** |
| Apple | Yes for all, but comment on 3) and 5) | On 3), we want to clarify that “UE selection” in two-side model has special requirement on NW to be aware of UE vendor information. According to RAN1 discussion on Direction A, the model training for UE-part model is performed per UE vendor (rather than per UE). Thus, NW needs to know UE vendor information due to below reasons:   1. If NW blindly selects UE without awareness of UE vendor info, it may result in some UE vendors can’t get sufficient data/parameters to complete its model training. 2. Risk of NW data leakage: For example, NW vendor A can only collaborate UE vendor A, but can’t achieve agreement on two-side model training with UE vendor B. Then, if NW vendor A blindly selects UE form UE vendor B and share data with it, it will lead to expose of NW proprietary information.   Thus, we suggest below modification:  3) UE selection **from one specific UE vendor** for being reception of dataset/model parameter    On 5), we agree to consider proprietary information of NW vendor. However, we think the concern is only on OTA transmission of parameter/model:   1. NW data leakage due to UE selection: NW vendor may blindly select UE of wrong UE vendor and share with it, as we mentioned in comment to UE selection. 2. OTA transmission on dataset/parameter from gNB to UE is not secure enough (E.g. OTA decoded by malicious UE nearby).   On non-OTA approach, we don’t see any security issue. Even if UE training entity is outside of 3GPP, SA2 had introduced Network Exposure Function (NEF) to securely expose the services and capabilities of 3GPP to external applications (e.g. UE-side server performs as AF and communicates with 3GPP network via NEF exposure). Meanwhile, we think such security concern is out of RAN2 scope and should be studied by SA2. Thus, we suggest:  5) proprietary information of NW topology and NW vendor should be respected and not disclosed **in air interface**. |
| vivo | See comments | These principles seem not critical for further discussion or solution design |
| Lenovo | In general OK |  |
| Huawei, HiSilicon | **For OTA:** not ok to 1), 2) and 5)  **For non-OTA:** add new principles | Firstly, for OTA approach, we have the following comments regarding the above principles:  for 1) and 2), we do not think they are essential due to relaxed latency for transmission. In other words, we believe that proper NW implementation can achieve a good balance between efficiency of dataset/parameter delivery and these impacts. We suggest to remove both principles.  for 3 and 4), non-OTA approach should follow similar principles as well.  for 5), for now RAN1 has not agreed on these information (e.g. information of NW topology and NW vendor), so the requirements are unclear to us. We suggest to remove this principle  **In summary, we suggest to remove 1), 2) and 5) for OTA approach.**  Secondly, for non-OTA approach, we have the following suggestions on new principles:  **a) UE-side server selection for transfer of dataset/parameter, e.g. how MNO co-ordinates with UE-side server**  **b) E2E reliability of dataset/model parameter transfer**  **c) Proprietary information of dataset/parameter should be respected and not disclosed**  Reason for adding a): how 3GPP NW node selects and co-ordinates with UE-side server needs to be discussed.  Reason for adding b): for non-OTA approach, the workflow may involve some NW entities, and how to guarantee E2E reliability needs to be discussed.  Reason for adding c): for inter-vendor case (e.g. NG-RAN/differen CN nodes are from different vendors), how to protect proprietary information of dataset/parameter and not disclose them across vendors need to be discussed. |
| ZTE | Only yes for 5 | I think 1,2,3,4 can be considered in normal phase |
| Qualcomm | Yes for all | We are aiming to develop a workable solution, therefore, concerns from UE vendors, infra vendors, and MNO need to be addressed.  For 1) relaxed latency does not implies that we should waste the air interface resources. Therefore, a solution for dataset / parameter transfer should minimize air interface wastage.  For 2) UE impact should be considered, as UE should not be kept in CONNECTED state just for the transfer of dataset / parameter (if UE does not have user traffic to transfer). Similarly, UE memory / power needs to be considered (I believe relaxed latency does not imply that NW will send a few bytes now and few bytes later after hours/days. It implies that if NW starts sending dataset / parameter, then UE has to allocate memory for storing received data until it is sent to the server).  For 3) we are okay with Apple change.  For 5) We have the same understanding as Apple. In non-OTA approach, CN procedures ensures that only an authorized AF can request the data exposed by CN functions. |
| OPPO | Yes for 5) | Compared to A1-A5, 1) 2) 3) principles are not common to OTA and non-OTA, my understanding is that dataset/model parameter transfer principle should be common enough to all the potential solution, in this sense, adding 1) 2) 3) may obviously be the burden for OTA solution which is really unfair guidance for the subsequent analysis.  For 4), I don’t know how RAN2 can evaluate this. It’s always true to consider reliability, but also means nothing, i.e. nobody knows what it is for now and how it impacts solution, so it’s too early to consider reliability.  For 5), this may impact how we design the solution details considering this is for cross vendor collaboration. |
| Ericsson | Yes to all | All these principles should be considered in the study.  Related to 5), we agree that from RAN2 point of view, this is mainly for the OTA solution. However, this is a general principle that should be applicable in general, irrespective of whether the solution will be OTA or non-OTA, i.e. the RAN-side data should not be disclosed without RAN involvement, irrespective of whether the entity collecting the NW-side dataset/model parameters is the UE (as in the OTA solution) or another NW entity (e.g. the OAM or CN in the non-OTA solution). |
| Nokia | **OTA**  No: 1, 2  Yes: 3, 4, 5  **Non-OTA**  No: 1, 2, 3  Yes: 4, 5 | **OTA**  We agree with Huawei that **1), 2)** are not necessary for OTA.  We think that **3)** is important, but we aren’t sure how that could be accomplished since the gNB does not have this information. **4)** We agree with this wording. At some layer, which might not be RRC, we need reliable transfer.  **5)** It isn’t clear how to accomplish this given the OTA diagram in the previous question. If the UE can send to a server over the internet, how do we guarantee that proprietary information will not be disclosed?  **Non-OTA**  **1)**, **2)**, and **3)** are irrelevant to Non-OTA |
| CATT | 1) 2) No  3) 4) 5) Yes | 1 and 2 can be considered when analysing different options, but they are not principles. |
| Mediatek | Yes for all | Both 1) and 2) are very reasonable principles to evaluate the OTA and non-OTA approaches.  For 1), scalability needs to be considered. The OTA approach should be scalable to handle an increasing number of UEs and larger datasets as AI models grow in complexity, ensuring that the network can support large-scale deployments without performance degradation. From a scalability point of view, relaxed latency does not necessarily mean that the impact on radio resources is negligible.  For 3), I have some sympathy for Apple's proposal. However, I am not sure whether it reflects a common understanding or specific assumptions in the RAN1 discussion. Therefore, the selection of UEs from one specific UE vendor needs to be confirmed by RAN1.  For 5), respecting the proprietary information of network topology and network vendors should be a general principle for both OTA and non-OTA approaches. RAN2 cannot evaluate which method is easier to protect proprietary information. Therefore, I suggest keeping 5) as it is. |
| Futurewei | Yes to 5), comments to others. | It seems only 5) is a principle while others are not; those are aspects that may be considered but no statements/principles have been made. |
| LGE | Yes | We generally share the same view with Qualcomm's intention in raising these considerations  For 1) and 2), we think those can be considered for super-sized data information. We share the same view with Qualcomm that even if the latency requirement is relaxed, air interface wastage by transferring dataset/parameter should be minimized.  For 3),4, and 5), we think those can be considered for two-sided model in the study. |
| Samsung | Yes but please see comments | We agree in general with all of them, but would like to note that only 5) and 4) give measurable guidelines. For 3) in particular it is not clear what is it that we are trying to optimize, and/or whether this is really what RAN1 asked us to do. In our understanding, UE will request providing data set/model parameters and therefore, we are not sure if there is a need for gNB to select UEs. |

## non-OTA solutions

From RAN2 point of view, RAN2 identifies following candidate solutions for non-OTA solution ‘NW dataset/model parameters collection entity -> UE side OTT server/UE training entity (inside/outside MNO)’, but no RAN2 impact is expected. Furthermore, considering RAN3, SA2, SA5 are impacted by below solutions, feasibility analysis of Non-OTA Option 1/2/3 is required to be evaluated by RAN3, SA2, and SA5. Note that it does not preclude RAN3/SA2/SA5 to identify other candidate solutions beyond options listed below.

|  |  |  |
| --- | --- | --- |
| **Option** | **Impacted WG** | **Specification impact/Implementation impact** |
| 1. OAM -> UE-side OTT server or UE-side training entity (inside/outside MNO) | SA5 | Up to SA5  (any intermediate node between OAM and UE-side OTT server is up to SA5; CN involvement if needed is up to SA2/SA5 discussion) |
| 1. CN -> UE-side OTT server or UE-side training entity (inside/outside MNO) | SA2 | Up to SA2  (any intermediate node between CN and UE-side OTT server is up to SA2) |
| 1. gNB -> OAM/CN -> UE-side OTT server or UE-side training entity (inside/outside MNO) | RAN3, SA2, SA5 | Up to RAN3, SA2, SA5  (any intermediate node between gNB/OAM, OAM/UE-side OTT server, CN/UE-side OTT server is up to RAN3/SA2/SA5) |

#### Q3-3: Do you think above non-OTA solutions can be considered as feasible solutions from RAN2 point of view, but with no RAN2 impact? Further feasibility analysis of non-OTA solution 1/2/3 is required to be evaluated by RAN3, SA2, and SA5. It does not preclude RAN3/SA2/SA5 to identify other candidate solutions beyond options listed below.

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **If no, please provide why it’s not feasible.** |
| Apple | Yes with comments | We see no technique showstopper from RAN2 perspective. And if any concern, it should be addressed by RAN3/SA2/SA5 directly because they fall into their expertise as captured in the table.  To avoid confusion to SA2/SA5/RAN3, we think it is necessary to clarify the difference between 1)/2) and 3). Our understanding is:   * In 1), parameter/dataset collection entity is CN. * In 2), parameter/dataset collection entity is OAM. * In 3), parameter/dataset collection entity is gNB.   We suggest to capturing the clarification in the table. |
| vivo | No, with comments | The feasibility of non-OTA solution is up to above other WGs, thus ‘non-OTA solutions can be considered as feasible solutions from RAN2 point of view’ should be removed. From R2 perspective, we can only say: “Feasibility analysis of non-OTA solution 1/2/3 is required to be evaluated by RAN3, SA2, and SA5. It does not preclude RAN3/SA2/SA5 to identify other candidate solutions beyond options listed below.” |
| Lenovo | Yes | We are also fine with the clarification suggested by Apple |
| Huawei, HiSilicon | No, with comments | For basic principles, here are our views:  A1: it may have impacts as NW may need to exchange lots of sets of dataset/parameters  A2: not relevant  A3: ok  A4: ok  A5: ok  For new principles (we proposed in Q3-2), here are our views:  **a) UE-side server selection for transfer of dataset/parameter**  How 3GPP NW node selects and co-ordinates with a UE-side server for transfer of dataset/parameter are unclear. In addition, how to make sure that the selected UE-side server can only get UE vendor-specific dataset/parameter is unclear. There may be a need for 3GPP NW to differentiate between UE vendors when generating dataset/parameter.  **b) E2E reliability of dataset/model parameter transfer**  For non-OTA approach, there may be interaction inside 3GPP NW nodes and interaction between 3GPP NW node and UE-side server, and how to guarantee E2E reliability is unclear.  **c) Proprietary information of dataset/parameter should be respected and not disclosed**  If one NW vendor generates dataset/parameter, and then passes them to UE-side server via CN/OAM. There is a risk that the proprietary information of dataset/parameter may be exposed to a second NW vendor without the original NW vendor's knowledge, which should be checked and discussed. |
| ZTE | Yes for 1, 2  No for 3 | In our understanding ,solution 3 is not an independent solution basically. It is just a whole picture of solution 1 and solution 2.  3 can be a note for non-OTA solution:  Note: It is up to RAN3, SA2, SA5 to discuss how and when gNB transfer the dataset or model parameter to OAM/CN. |
| Qualcomm | Yes | We are also fine with the clarification suggested by Apple |
| OPPO | No, with comments | We have the similar feeling with vivo, it’s hard to say RAN2 thinks it’s feasible for these non-OTA solutions. If people think this falls into the scope of RAN3/SA2/SA5, we should check with them just as we did for UE side data collection Option2/3, the feasibility is evaluated by SA2/SA5 triggered via LS, so we don’t know why we have different treatment for dataset/model parameter transfer, in this sense, we tend to propose:  **“Feasibility analysis of non-OTA solution 1/2/3 is required to be evaluated by RAN3/SA2/SA5. It does not preclude RAN3/SA2/SA5 to identify other candidate solutions beyond options listed below.”** |
| Ericsson | Yes, but the 3 solutions are now overlapping, so we need to distinguish what are the use cases we have in mind. We can simplify in:   1. For the OAM centric:    1. OAM -> UE-side OTT server or UE-side training entity (inside/outside MNO) 2. For the gNB-centric:    1. OAM/CN -> UE-side OTT server or UE-side training entity (inside/outside MNO) | All the three options are ok in principle, but considering that the link “gnB->CN/OAM” is already covered by P2 above, we suggest removing the gNB and distinguish the cases of OAM-centric and gNB centric.  For the OAM centric, the dataset/parameters are generated by the OAM and shared with the training entity (possibly via CN, as stated in the 3rd column).  Whereas for the gNB-centric, the dataset/parameters are generated by the gNB and shared with the OAM/CN that are the NW-side dataset collection entities in this case. training entity  Related to Apple´s comment, we are fine to clarify which is the collection entity. However, we believe that the proposal from Apple is not correct:   * In 1), parameter/dataset collection entity is OAM.   + As said above, this is the case where for example of OAM-centric data collection, and data are passed directly to the server or to another CN node (as mentioned in the 3rd column). * In 2), parameter/dataset collection entity is CN.   + This is the case where for example the dataset goes from the gNB to the CN in case of gNB centric training or from OAM to CN in case of OAM centric training. Hence this case can be removed since already covered in the option 1) above and 3) below. * In 3), parameter/dataset collection entity is OAM/CN.   + For example in the case of gNB-centric data collection, the gNB generates the dataset/model parameters and pass it to the OAM/CN (which are in this case the NW-side dataset/model parameters collection entity).   Relate to other comments above, please note that RAN1 explicitly “ask RAN2’s feedback on the **feasibility of a standardized signaling**, via over-the-air and/or other approaches”. Hence, we believe that RAN2 should assess the feasibility (from our point of view) of all the possible solutions, and request other WGs to further assess feasibility (as mentioned in this proposal). |
| Nokia | No | Agree with Vivo. We can leave it to the other WGs to determine feasibility. We do not have the expertise or authority to do so for the non-OTA approaches. |
| CATT | Yes |  |
| Mediatek | Yes for 1, 2  No for 3 | We agree with ZTE that the path from gNB to OAM/CN is common for option 1 and 2, which is unavoidable in the process of parameter/dataset sharing from the network to the UE training entity in the non-OTA approach.  We should avoid defining different variations with overlapping elements.  Similar to data collection, we only need to consider options 1 and 2 in the table above. |
| Futurewei | Yes. | Agree in general. |
| LGE | Yes | We are also fine with the clarification suggested by Apple |
| Samsung | Please see comment | OK in principle but as the rapporteurs state, this should not preclude other WGs from identifying other solutions, and/or (in our view) perhaps even rejecting some of the solution options (paths) that we list. We therefore do not see the usefulness of this info or its intended use? We could simply say (as vivo suggests) that non-OTA is out of RAN2 scope.  Regarding rapporteurs’ comments (related to our Phase-1 input) which compares this analysis to the UE-side data collection solution analysis, and claims RAN2 remit in both, in our view in case of UE-side data collection, data is transferred from the UE to NW entities (CN, OAM, UE server) and RAN2 has some expertise/remit there from protocol point of view. However, from NW entity to UE server over non-OTA has nothing to do with RAN protocols. |

## OTA solution (NW dataset/model parameter entity -> UE)

After phase 1 discussion and companies’ feedback, rapporteur further summarized below feasibility/complexity analysis for each candidate solution:

OTA solution 1a: gNB -> UE via CP Feasibility Analysis

As mentioned by some companies during phase 1 that analysis in TR38.843 could be reused. Two companies raised the concern of simply scaling up RRC segment number is not feasible.

|  |  |  |
| --- | --- | --- |
| **Solution** | **Views on whether solution 1a as a candidate solution** | **Enhancements to be considered for feasible solutions/Reason of not feasible** |
| 1a | **Yes (12/14)**: ZTE, Apple, HW, vivo, QC, Oppo, MTK, SS, Ericsson, LG, CMCC, FW | * **Extend existing supported RRC signaling size (same spec impact of solution 1a in TR38.843)**: ZTE, Apple, HW, Vivo, QC, Oppo, MTK, SS, LG, CMCC |
| **No (2/14):** Lenovo, Nokia | * **not scalable for size up to hundreds of MB:** Lenovo, Nokia (shouldn’t extend when there’s a viable non-OTA method available) |

Rapporteurs understand the concern raised that scaling RRC segment number to support 225MB might be challenging. On the other hand, there might be other segmentation methods can be considered to reduce segmentation challenges in RRC, as mentioned in [3] and [9].

According to the discussion in Q1-1, rapporteurs further add a new column indicating whether the previous analysis can be applicable to dataset/model parameter transfer.

|  |  |  |
| --- | --- | --- |
| **Current status and Gaps of Solution 1a in TR38.843** | **Potential RAN specification impact of Solution 1a in TR38.843** | **Applicable to dataset/model parameter transfer?** |
| Model size >45kBytes is not supported based on existing number of RRC segments | Extension of the number of RRC segments is required to support models larger than 45kBytes | **Yes (A1-Size),** but other segmentation method is not precluded (as mentioned by [3][9]). The detailed segmentation solutions for this hundreds of MB dataset/model parameter size can be further discussed during WI. Rapporteur suggests to revise it as below:  **Segmentation to support hundreds of MB dataset/model parameter transfer** |
| Transmission is restarted upon mobility | - Requires service continuity support for SRBs with segmentations.  - Xn/NGAP enhancement(s) for model transfer/delivery continuity | **Yes with change (A2- Continuity):**  - Requires service continuity support for SRBs with segmentations during UE mobility  - Xn/NGAP enhancement(s) for dataset/model parameter transfer continuity |
| Management and interaction between UE and gNB is not supported | Requires management and interaction between UE and gNB (e.g., model identification, model transfer completion indication, etc.) when model management at gNB | Not applicable since it’s not model transfer |
| Procedure latency depends on model size and SRB priority | Impact on SRB in DL, e.g., a new SRB with configurable priority, etc. | **Yes (A4- Latency)** |

Rapporteurs observe that the above specification impact seems manageable from RAN2 point of view.

Furthermore, as summarized in Q1-1, A3-Controllaiblity and A5-Visibility also need to be supported in solution 1a, where A3 is naturally supported by RRC signaling and A5 depends on the standardized data format and model structure (which is up to RAN1).

Based on above analysis, following specification impact can be considered to achieve OTA solution 1a feasibility from RAN2 point of view:

* Segmentation to support hundreds of MB dataset/model parameter transfer
* Requires service continuity support for SRBs with segmentations during UE mobility
* Xn/NGAP enhancement(s) for dataset/model parameter transfer continuity
* Impact on SRB in DL, e.g., a new SRB with configurable priority, etc.

#### Q3-4: Do you think candidate solution OTA solution 1a (i.e. gNB -> UE via CP) is feasible with manageable specification impact listed above?

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Yes/No** | **If yes, how it is feasible/not complex?**  **(Please address companies’ concern during phase 1)** | **If no, why it’s not feasible/complex?** |
| Apple | No |  | First, as UE/chipset vendor, we think it is not feasible for 5G UE to have RRC buffer >200MByte (e.g. at least for data/parameter segments reassemble):   * Please note that section 4.2.14 of TS 38.306 clearly captured below: “*The RRC buffer size is defined as* ***the maximum*** *overall RRC configuration* ***size that******the UE is required to store. The RRC buffer size is 45Kbytes****.”* This limit was specified in Rel-15, and UE vendors had designed their memory access and hardware accelerators for ciphering based on it. Supporting >200Mbyte RRC buffer would require major re-design of the user plane hardware architecture, which is not realistic in late 5G. * RRC buffer utilizes RAM memory which is a scarce resource rather than UE internal storage. And limited RAM memory for baseband needs to share with RF buffer (e.g. for analog beamforming), L2 buffer (e.g. for RLC/PDCP record/reassemble), and L1 buffer (e.g. for massive MIMO processing). We think it is quite challenging for 5G UE to have 200Mbyte RRC buffer.   Secondly, we understand the proposed other segmentation solutions (e.g. in [3][9]) are not legacy RRC segmentation specified in RAN2, but segmentation in a new layer above RRC (e.g. AI/ML layer). Note in Rel-16, RAN2 specify **RRC segmentation is performed** **after ASN.1 coding**. So, these proposals essentially introduce a new SRB protocol stack in late 5G because RAN2 requires to specify how to generate AI/ML PDU from AI/ML SDU and design its PDU header (at least include segmentation info).  Finally, we don’t think the spec work is manageable. At least below RAN2 spec changes are expected:   1. New SRB protocol stack with a new specified AI/ML layer above RRC, including how to generate AI/ML PDU, PDU header and status report PDU (for continuity). 2. RRC segments continuity: current 38.331 specify the UE discards previous RRC segments in HO. 3. SRB recovery: current 38.322 only support data recovery in DRB. 4. New SRB design: current 38.300 specify all segments of an RRC message are transmitted before sending another RRC message. 5. How to select UE only from one specific UE vendor.   We have to emphasize that the long list is conflicted with sprit of RANP#107 (e.g. critical to have reasonable amount of TUs accommodating necessary 5G-Advanced items to address commercial needs) |
| vivo | Yes, comments | To address companies’ concerns, OTA solution can start with model parameter transfer with smaller size.  For model parameter the size is less than hundred MB. Please note that from R1 Ls, R2 can just assume a model parameter size of 11.6MB, not hundreds of MB. As, considering hundreds of MB is not appropriate for model parameters size. So, we suggest to revise the 1st bullet as:   * **Segmentation to support hundreds of MB dataset~~/model parameter~~ transfer, and smaller number of segmentations to support tens of MB model parameters transfer** |  |
| Lenovo | No |  | The first bullet “Segmentation to support hundreds of MB dataset/model parameter transfer” is the fundamental basis for the option “gNB->UE via CP”. It implies quite big change to the existing use/capability of RRC buffer at UE as also pointed out by Apple. Comparatively, the CN/OAM -> gNB -> UE via UP is a more feasible approach, if we really want to support OTA.  On top of that, of course, handling the UE mobility during “days/weeks” will add additional complex.  We are not against listing “gNB->UE via CP” and possible spec impact as listed by the rapporteur in the TR to show RAN2 analysis. **But it is difficult to say the solution is “feasible”, indicating it’s practical in real deployment.** |
| Huawei, HiSilicon | Yes, with comments | Firstly, as mentioned in A4 - Latency, the transmission of dataset/parameter is different from normal user traffic data. So NW should be able to have flexible means to transfer them to UE via Uu, e.g. NW can select proper UEs at proper time to do the dataset/papater transfer, and the integrity/efficiency can be guaranteed.  Secondly, if data size is small, there should be no problems as NW can transfer dataset/parameter to one UE via an efficient way.  If data size is large, we think OTA approach is feasible, due to the following reasons:  (a) NW can split the data into some pieces (either in RRC layer or higher layer), and then these pieces can be transferred to one UE or multiple UEs. Since there are ids to guarantee the integrity, UE-side server can merge them into a whole one. If some pieces have not got by UE-side, UE can request them from NW. Details can be left to later phase if possible  (b) For service continuity and Xn impacts, as we mentioned for (a), there are no problems because if the UE just gets some pieces, the UE could be able to request left pieces from NW, or there may be some signalling impacts and overhead if NW wants to continue the data transmission  (c) for impact on SRB in DL, since A4 - Latency is assumed here, NW can select proper UEs at proper time so that the data transmission has minimal impacts to other UEs. A low-priority SRB can be also considered for data transmission in Uu interface |  |
| ZTE | Yes | I understand the main concern is about the significant gap between the model/data set size and maximum DL RRC signaling size.  As rapporteur point out, in addition to directly increasing the segment number of DL RRC message, other segmentation method has been mentioned in [1] and [9], which is, one UE can be responsible for a small part of dataset/model parameter, and then forward to UE side server for the further processing. |  |
| Qualcomm | Feasible; at significant specification impact and implementation complexity. |  | Apart from the reasons mentioned by Apple and Lenovo, there are significant specification impact associated with   * UE selection for dataset/parameter transfer from a UE vendor, * E2E reliability of dataset/model parameter transfer (e.g., handling during UE mobility, RRC state transitions). |
| OPPO | Yes | On one hand, one UE can be responsible for a small part of dataset/model parameter, this gives the way to limit the segmentations.  On the other hand, dataset and/or parameter sharing size could range from tens of KBs to hundreds of MBs, for the small size, RRC still can be the candidate solution while for large size, gNB can try to distribute the burden of each UE, which is also feasible. |  |
| Ericsson | No | The specification cost and the standardization efforts to specify this functionality is huge | We agree with the previous observations from skeptical companies. We would like to further capture the following issues:   * RRC signalling scalability: The maximum RRC and SIB message size cannot carry the large amount of data expected for dataset/model parameter delivery over-the-air. Further, RRC protocol requires that all segments of an RRC message are transmitted before sending another RRC message. * Over-the-Air Identification: As mentioned by some companies above, a mechanism is needed to uniquely identify a dataset/model parameter and its corresponding partition IDs (if the dataset transmission is shared by different gNBs) across different vendors. * Increased Complexity at UE/gNB: Both the NW and UE would face additional complexity in managing which parts of the dataset/parameters should be transmitted through specific gNBs and delivered to specific UEs in the network. * Uu overhead: 1)The NW should first collect data over the UL, in order to generate the dataset/model parameters. 2)Then the generated dataset/model parameters should be sent back to the UE in the DL, and finally 3) the UE sends them again over the UL to the training entity.  Compared with non-OTA approach this means that impact over the Uu air interface is much higher, given that for non-OTA only the step 1) is needed. |
| Nokia | No |  | We **agree with Apple and Lenovo** on their respective feasibility analyses regarding increasing the maximum number of segments and in creating a new type of SRB.  Furthermore, we do not see any value in this large amount of specification work since an alternative, non-OTA solution, which would meet all the requirements without burdening the air interface, is also possible.  We disagree with vivo’s approach as splits the dataset and parameter set mechanisms instead of unifying them. There is no identified need to transfer small datasets and parameter sets over RRC while transferring larger ones through non-OTA means. The latency requirement is low.  To support the approach proposed by Huawei, we would still need to build a protocol atop RRC to handle retransmissions and to track the status of the overall transmission – removing duplicate segments and identifying missing segments. We do not think this is trivial, and since protocols such as TCP/IP, which can be used in the non-OTA solution, already supports these features, we do not see the benefit of recreating this over the uu CP interface.  The transfers would not be done by many UEs. What we are discussing here is committing significant effort to adding a feature that a small minority of UEs would likely use given that these dataset and parameter set transfers shouldn’t be happening often. |
| CATT | Yes with comments | The split segmentation options discussed for UE radio capability optimization can be considered. Since the dataset/model parameter size provided by RAN1 needs thousands of segments which greatly exceeds RRC buffer size defined for current UE capability, “Hard” split RRC segmentation without multiple decode is applicable for AI model transfer if the RRC buffer size can be expanded.  Since the latency requirement is relaxed, the UE can decode the complete information after assembling all parts, no matter a new SRB is used for the dataset/model parameter transfer.  For service continuity, the enhancement of Xn/NGAP is needed, but we do not think it is a complicated enhancement as it’s based on the cable network. |  |
| Mediatek | No |  | I agree with the technical reasons mentioned by Apple and Lenovo. If we aim to develop a workable and practical solution, supporting thousands of segments with an extremely long sharing process is not viable, especially when we want to commercialize and scale AI over wireless.  The CP is originally designed for compact signaling and control messages, not for large data transfers. Furthermore, overloading CP with large datasets would disrupt its core functions, leading to potential service degradation and reliability issues.  For AI over wireless to be commercially viable and scalable, the data transfer process must be efficient and cost-effective. The OTA solution 1a is unaffordable for the network |
| Futurewei | Yes. | While not the best solution, we think it is doable. |  |
| LGE | No |  | Due to the current size limitation of RRC signalling (with a maximum constraint of 9000 bytes), segmentation must be supported to deliver the super-sized dataset/model parameters. Transferring large amounts of data requires multiple RRC transmissions, which leads to increased signalling overhead. Additionally, the transfer of numerous segments heightens the risk of missing data segments during UE mobility, potentially compromising service continuity.  Further feasibility check is needed by other WGs. |
| Samsung | Yes with comments | Regarding concerns from Nokia and Lenovo on RRC scalability – while we understand the general concern, we note that the TR already took into account the fact that that the model size could be significant, and possibly comparable with what we are now discussing for data set/model parameter transfer. Therefore we’re not sure if the concern here is the feasibility of the required change to RRC, or its necessity?  In our view OTA is necessary – as we shared in Phase 1, we are not at this point sure that direct server-to-server transfer (i.e. non-OTA) is even possible in all jurisdictions / from legal point of view. Therefore we think there is a necessity for OTA. If opponents of the CP approach feel that there is a size threshold where required RRC enhancements are no longer viable, it would be good to know it. |  |

OTA solution 2: CN -> UE via gNB Feasibility Analysis

|  |  |  |
| --- | --- | --- |
| **Option** | **Views on whether can be considered as a candidate solution** | **Enhancements to be considered for feasible solutions/Reason of not feasible** |
| CN -> UE via gNB | **Yes (6/14):** ZTE, vivo, Lenovo, Oppo, SS, FW (yes for model parameter) | * **only consider when dataset/model parameter is generated NW training entity (CN)**: SS |
| **No (8/14):** Apple, HW, QC, MTK, Ericsson, Nokia, LG, CMCC | * **dataset/model parameter is transparent to gNB:** Apple * **no benefit compared to non-OTA approach as dataset/model parameter is generated by gNB**: HW, QC, MTK * **haven’t study feasibility of CN as NW training entity:** Ericsson * **leave it to other WGs:** LG |

First of all, as analysed in TR 38.832 Table 7.3.1.4-2, Table 7.3.1.4-4, and Table 7.3.1.4-6, it seems all solutions above can meet principles for dataset/model parameter transfer with certain specification impacts.

However, when CN is the NW dataset/model parameter collection entity, some companies commented that this solution is complex and no benefit because the collected dataset/model parameters need to be routed via different network entities and then transfer to UE (as shown in Alternative 2 figure considering dash line is valid). Therefore, it seems there’s not much benefit compared to non-OTA approaches. Technically speaking, they are feasible solutions, but from complexity point of view, it may introduce extra overhead. Especially when dataset/model parameter generated by gNB, data needs to be transmitted twice over gNB.

#### Q3-5: Do you agree that OTA solution 2 (i.e. CN -> UE via gNB) is feasible but complex? The feasibility of CN -> UE via gNB needs to be further confirmed by RAN3 and SA2.

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Yes/No** | **If yes, how it is feasible/complex?** | **If no, why it is not feasible/ not complex?**  **(Please address companies’ concern during phase 1)** |
| Apple | No |  | 1. It is not feasible: as long as the data/parameter is still transferred via SRB, we have the same issues mentioned in Q3-4. 2. It is not necessary: we agree other companies that this solution has no benefit over non-OTA solution as data/parameter are generated by gNB. In RAN2, we generally don’t specify two solutions for same issue. |
| vivo | Yes | Feasible from R2 perspective. Further analysis can be left to RAN3 and SA2. |  |
| Lenovo | Yes, with comment | We agree there is no technical benefit comparing to non-OTA approach considering the repetitive transmission in DL and UL, and this is the same for all Solution 1a, 2, 3.  If we really want to support OTA approach, comparing to Solution 1a, solution 2 and 3 can be over user plane, which are more feasible from UE point of view.  We agree the feasibility needs to be further confirmed by RAN3 and SA2. |  |
| Huawei, HiSilicon | See comments | Firstly, we think only gNB can generate dataset/parameter.  Secondly, OTA solution 2 has pros and cons, and TR 38.843 has provided some analysis, i.e. analysis of Solution 2a and 2b in section 7.2.1.4.  In addition, we also observe the following issues (using new principles as proposed in Q3-2):  **b) E2E reliability of dataset/model parameter transfer**  For OTA solution 2, there may be interaction between gNB and CN, and interaction between CN and UE, and how to guarantee E2E reliability is unclear.  **c) proprietary information of dataset/parameter should be respected and not disclosed**  If one NW vendor generates dataset/parameter, and then passes them to CN. There is a risk that the proprietar information of dataset/parameter may be exposed to a second NW vendor without the original NW vendor's knowledge, which should be checked and discussed. |  |
| ZTE | See comments | Before discussing the feasibility and complexity, we need to discuss the necessity in advance. Whether to support the dataset/model sharing from CN to UE mainly depends on where the model parameter or data set is stored (e.g. gNB, CN or OAM), in other words, where the model is trained at NW side, according to the TR,for CSI feedback case, the training entity can be CN but it cannot be studied in RAN2, please see below:  ------------------ From TR ----------------  -Model Training:  oFor 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.  ------------------ From TR ----------------  In this sense, as TR said, the feasibility discussion is out of RAN2 scope, let along the complexity discussion. In this sense, we suggest not to analyze the feasibility at the current phase, instead  We can add a similar note for this solution:  *Note: RAN2 identified the CN may need to share the dataset/model parameter with UE via gNB, the feasibility and complexity cannot be studied in RAN2 as the path between CN and gNB is out of RAN2 scope.* |  |
| Qualcomm | No |  | If the dataset / parameter is already available in CN, then OTA solution 2 is not desirable. |
| OPPO | See comments | We don’t think RAN2 can preclude this direction by RAN2 alone, anyway the feasibility of CN -> UE via gNB needs to be further confirmed by RAN3 and SA2. |  |
| Ericsson | No |  | We agree with ZTE comment that in the topic of NW-side data collection, we never discussed how the CN can get the collected data from the gNB. Assuming that the CN is the dataset/model parameter collection entity (which is one of the non-OTA solutions outlined above), then we agree with QC, that it does not sounds reasonable that the dataset is sent back to the gNB for OTA delivery. |
| Nokia | No |  | We agree with Apple’s analysis. Also, when we discuss feasibility, we should consider technical feasibility as well as standardization feasibility. Considering the vast amount of work required (for potentially no benefit over non-OTA), we also see this as infeasible to standardize. |
| CATT | Yes | It is feasible since the CN can transfer dataset/model parameter via cable network. We do not see much complexity compared with solution1. |  |
| Mediatek | No |  | Agree with Qualcomm |
| Futurewei | Yes and No | Yes for model parameters. For model parameters, the UE is the target of the transfer. Also, it is possible that the model is stored and managed at the CN. | No for dataset transfer. For training dataset transfer, the target could be a UE-side training server. There is no point of sending the dataset to the UE then transfer it to the UE-side training server, unless the training is done at the UE. |
| LGE | Yes | Further feasibility check is needed by other WGs (RAN3/SA2). We don’t see any benefit with this solution comparing to non-OTA solution. |  |
| Samsung | Yes with comments | We agree it needs to be confirmed by RAN3 and SA2, but we see no need (or possibility) for RAN2 to ‘block’ it.  We further do not see the need to say “complex”. This scenario is just the opposite direction of Option 2 in UE-side data collection. And we do not say Option 2 is “complex”. Besides, RAN1 just asked about “feasibility”, so we do not need to mention “complexity”. |  |

OTA solution 3: OAM -> UE Feasibility Analysis

|  |  |  |
| --- | --- | --- |
| **Option** | **Views on whether can be considered as a candidate solution** | * **Enhancements to be considered for feasible solutions/Reason of not feasible** |
| OAM -> UE via gNB | **Yes (7/14)**: ZTE, vivo, Lenovo, Oppo, SS, Ericsson | * **only consider when dataset/model parameter is generated NW training entity (OAM)**: SS, Ericsson |
| **No (6/14):** Apple, HW, QC, MTK, Nokia, LG, CMCC | * **dataset/model parameter is transparent to gNB:** Apple * **no benefit compared to non-OTA approach as dataset/model parameter is generated by gNB:** HW, QC, MTK * **leave it to other WGs:** LG |

Similar as OTA solution 2, when OAM is the NW dataset/model parameter collection entity, it is feasible to meet principles to transfer dataset/model parameter from OAM to UE via gNB, by considering specification impact analysed in TR 38.843 Table 7.3.1.4-6. However, when OAM is the NW dataset/model parameter collection entity, the same analysis of overhead for OTA solution 2 also applies to OTA solution 3.

#### Q3-6: Do you agree that OTA solution 3 (i.e. OAM -> UE via gNB) is feasible but complex? The feasibility of OAM -> UE via gNB needs to be further confirmed by RAN3 and SA5.

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Yes/No** | **If yes, how it is feasible/complex?** | **If no, why it is not feasible/not complex?**  **(Please address companies’ concern during phase 1)** |
| Apple | No |  | 1. It is not feasible: as long as the data/parameter is still transferred via SRB, we have the same issues mentioned in Q3-4. 2. It is not necessary: we agree other companies that this solution has no benefit over non-OTA solution as data/parameter are generated by gNB. In RAN2, we generally don’t specify two solutions for same issue. |
| vivo | Yes | Feasible from R2 perspective. Further analysis can be left to RAN3 and SA5. |  |
| Lenovo | Yes with comment | We agree there is no technical benefit comparing to non-OTA approach considering the repetitive transmission in DL and UL, and this is the same for all Solution 1a, 2, 3.  If we really want to support OTA approach, comparing to Solution 1a, solution 2 and 3 can be over user plane, which are more feasible from UE point of view.  We agree the feasibility needs to be further confirmed by RAN3 and SA5. |  |
| Huawei, HiSilicon | See comments | Firstly, we think only gNB can generate dataset/parameter.  Secondly, OTA solution 3 has pros and cons, and TR 38.843 has provided some analysis, i.e. analysis of Solution 4b in section 7.2.1.4.  In addition, we also observe the following issues (using new principles as proposed in Q3-2):  **b) E2E reliability of dataset/model parameter transfer**  For OTA solution 3, there may be interaction between gNB and OAM, and interaction between OAM and gNB/UE, and how to guarantee E2E reliability is unclear.  **c) proprietary information of dataset/parameter should be respected and not disclosed**  If one NW vendor generates dataset/parameter, and then passes them to OAM. There is a risk that the proprietar information of dataset/parameter may be exposed to a second NW vendor without the original NW vendor's knowledge, which should be checked and discussed. |  |
| ZTE | See comments | As we said in Q3-5, before discussing the feasibility and complexity, we need to confirm the necessity, in TR 38.843:  ------------------ From TR -------------  -Model Training:  oFor 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.  ------------------ From TR -------------  For OTA solution, the necessity of supporting the path from OAM to UE via gNB is obvious. However, the evaluation of this solution feasibility is out of RAN2 scope let along the complexity. In this sense, we suggest not to give the feasibility and complexity in the current phase, instead we can add a note:  *Note: RAN2 identified the OAM may need to share the dataset/model parameter with UE via gNB, the feasibility and complexity cannot be studied in RAN2 as the path between OAM and gNB is out of RAN2 scope.* |  |
| Qualcomm | No |  | If the dataset / parameter is already available in OAM, then OTA solution 3 is not desirable. |
| OPPO | See comments | We don’t think RAN2 can preclude this direction by RAN2 alone, anyway the feasibility of OAM-> UE via gNB needs to be further confirmed by RAN3 and SA5. |  |
| Ericsson | Feasibility of Q3-6 is subject to feasibility of Q3-4 | The impact of OAM->gNB should be considered for the OAM-centric data collection, and studied by SA5.  Feasibility concerns for OTA delivery are the same as in Q3-4. | Even if we go for OTA, we need to consider the way the dataset generated by the OAM goes to the gNB, for the case of OAM-centric data collection. In such case, it is the OAM that generates the dataset/model parameters, and the way this data goes from the OAM to the gNB should be studied by SA5. We cannot simply ignore the OAM-centric framework.  Regarding the OTA impacts (once the dataset is available at the gNB), the same comments as in Q3-4 apply. |
| Nokia | No |  | We agree with Apple’s analysis. Also, when we discuss feasibility, we should consider technical feasibility as well as standardization feasibility. Considering the vast amount of work required (for potentially no benefit over non-OTA), we also see this as infeasible to standardize. |
| CATT | Yes | It is feasible since the CN can transfer dataset/model parameter via cable network. It is more complex compared with solution2 which needs enhancement in both RAN3 and SA5. |  |
| Mediatek | No |  | Agree with Apple and Qualcomm.  If the dataset/parameters are already available in OAM, the OAM should share them directly with the UE training entity. Relaying them via gNB and UE would consume double the resources in both the network and air interface, which is quite inefficient. |
| Futurewei | Yes and No | Yes for model parameters. For model parameters, the UE is the target of the transfer. Also, it is possible that the model is stored and managed at OAM. | No for dataset transfer. For training dataset transfer, the target could be a UE-side training server. There is no point of sending the dataset to the UE then transfer it to the UE-side training server, unless the training is done at the UE. |
| LGE | Yes | Further feasibility check is needed by other WGs (RAN3/SA2). We don’t see any benefit with this solution comparing to non-OTA solution. |  |
| Samsung | Yes with comments | We agree it needs to be confirmed by RAN3 and SA5, but we see no need (or possibility) for RAN2 to ‘block’ it.  We further do not see the need to say “complex”. This scenario is just the opposite direction of Option 2 in UE-side data collection. And we do not say Option 2 is “complex”. Besides, RAN1 just asked about “feasibility”, so we do not need to mention “complexity”. |  |

# 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