**3GPP TSG RAN2 #121 R2-23xxxxx**

**Athens, Greece, 27th Feb – 3rd Mar, 2023**

**Agenda Item:**  **XX.YY**

**Source: Huawei (email rapporteur)**

**Title:** **Report of [Post120][053][AIML18] model transfer delivery (Huawei)**

**Document for: Discussion and Decision**

# 1 Introduction

This is the email report of [Post120][053]:

* [Post120][053][AIML18] model transfer delivery (Huawei)

Scope: Long email discussion for next meeting on model transfer/delivery, to collect pros/cons, Can also collect comments on different architectural assumptions.

Intended outcome: Report

Deadline: Long (10th Feb, 10:00 UTC, 2023)

As indicated by the Chair, the inactive period is:

* Dec 23 – Jan 6 is an expected inactive period (for confirmation TSG RAN)
* Jan 23 – 27 is an inactive period (for confirmation TSG RAN)
* Also Weekends are inactive

In order for efficient discussions, it is suggested to have 2 phases:

**Phase 1:**

The deadline is 13th Jan, 10:00 UTC. The phase 1 summary will be provided by 16th Jan 10:00 UTC, and then companies can check it.

In this phase, it is suggested to discuss the terminologies “model transfer/delivery”, and also architectural assumptions. For architectural assumptions, there are some high level discussions on options and applicable use cases, and then the outcome of this part will be used for phase 2 discussion, e.g. phase 2 will focus on possible options.

**Phase 2:**

From 17th Jan to 10th Feb, 10:00 UTC. The phase 2 summary will be provided by 13th Feb 10:00 UTC, and then companies can check it. The final summary will be submitted by the submission deadline (likely 17th Feb).

Based on the outcome of Phase 1 discussion, for phase 2, it is suggested to discuss model transfer/delivery in Downlink and Uplink, i.e. whether to focus on model transfer/delivery in DL in this email discussion. For each option for CP/UP-based solutions, it is suggested to discuss principles and basic flows, and then discuss pros/cons.

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

|  |  |  |
| --- | --- | --- |
| **Company** | **Name** | **Email Address** |
| OPPO | Jiangsheng Fan | fanjiangsheng@oppo.com |
| Qualcomm | Rajeev Kumar | rkum@qti.qualcomm.com |
| vivo | Xiang Pan | panxiang@vivo.com |
| NEC | Xuelong Wang | xuelong.wang@emea.nec.com |
| Apple | Peng Cheng | pcheng24@apple.com |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# 2 Discussion

## 2.1 Phase 1

### 2.1.1 Discussion on terminologies

In RAN1, model transfer and model delivery have been defined, i.e. Table 1. In RAN2, companies are using the terminologies but there are different understandings regarding solution details.

**Table 1: Terminologies for AI/ML model transfer and AI/ML model delivery**

|  |  |
| --- | --- |
| AI/ML model transfer | Delivery of an AI/ML model over the air interface, either parameters of a model structure known at the receiving end or a new model with parameters. Delivery may contain a full model or a partial model. |
| AI/ML model delivery | A generic term referring to delivery of an AI/ML model from one entity to another entity in any manner.  Note: An entity could mean a network node/function (e.g., gNB, LMF, etc.), UE, proprietary server, etc. |

It is suggested to align the wording in this email discussion (not touching the concept):

- Option 1: Use the wording “model delivery” and it covers both model transfer and delivery

- Option 2: Use the wording “model transfer/delivery”

From the email rapporteur’s point of view, the main discussion on AI/ML model transfer/delivery is for “UE-sided model/UE-part model for two-sided model”, and RAN2 could study the procedures for possible options. If some procedures are out of RAN2 scope, RAN2 may have some initial discussions and then decide how to progress on them (e.g. RAN2 might check with other WGs later).

**Q1: Regarding how to use the terminology model transfer/delivery in this email discussion, which option is preferred?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Option 1/2** | **Comments** |
| OPPO | Option2 | As we know, both terminologies are defined by RAN1, model delivery is a more generic concept than model transfer, while model transfer focuses on delivery of an AI/ML model over the air interface. We may select one of the terminologies during normative work if recommended, but for now, nearly all things are open for model delivery/transfer. On one hand, no additional gain RAN2 will get to differentiate these two terminologies at this early stage especially considering RAN1 is also debating something for this; on the other hand, without debating this, we may have a more efficient discussion for this email, so the safer and easier way is to keep both terminologies for now in our discussion. |
| Qualcomm | Option2 | Option 2 is more generic and avoid unnecessary confusion. |
| vivo | Option 2 | For Model transfer, UE will be involved as the model is delivered via the air interface. While model delivery is a generic term, which includes model delivery between NW entities (e.g., RAN node and CN), and the model can be delivered via other wireless technologies, e.g., WiFi.  As UE is involved in all the architectural options in section 2.1.2, this email is actually discussing the solutions of model transfer.  Therefore, the model transfer shall be reflected in the conclusion of this email. |
| NEC | Option 2 | From terminology perspective, we have a slight preference to just align with what RAN1 is using.  Meanwhile, we are aware that different WGs may use different terminology for model transfer/delivery. In additions to the RAN1 definition, in TR37.817, RAN3 uses the term “Model Deployment/Update” to represent “deploy a trained, validated, and tested AI/ML model to the Model Inference function or to deliver an updated model to the Model Inference function.”.  From the air interface perspective, we sharing the understanding that model transfer may be adopted for this purpose. |
| Apple | Prefer Option 2, Option 1 is acceptable | First, we don't think it is an essential issue but just terminology issue. Although we understand "model delivery" is a more general concept to include both "model transfer" and "delivery", we slightly prefer to use "model delivery/transfer" to avoid any possible misalignment with RAN1. That is because both terminologies are defined in RAN1. Although we understand RAN1 actually only use "model delivery" in their discussion now, we are not sure whether each company in RAN2 has the same understanding which may cause unnecessary discussion in RAN2. So we prefer RAN2 can just use model delivery / model transfer for study and wait RAN1's further input. |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

### 2.1.2 Architectural assumptions

For this email discussion, one target is to collect comments on different architectural assumptions. This section is to have some high level discussions to check whether all combinations are agreeable for RAN2 study or not. After this discussion, phase 2 can discuss details and pros/cons of each possible option.

At RAN2#120, the Chair made the following statement, which has been considered in this email discussion.

*Chair: It is allowed to discuss/determine that functionality can be done outside 3GPP system scope, i.e. OTT server. NO agreement for now on the specifics due to long discussion.*

Based on the contributions at RAN2#120, the following options on architectural assumptions are provided:

* Option 1: Model transfer/delivery between UE and gNB. For this option, CP and UP solutions can be studied
* Option 2: Model transfer/delivery between UE and CN. For this option, CP and UP solutions can be studied
* Option 3: Model transfer/delivery between UE and LMF. For this option, CP solution can be studied
* Option 4: Model transfer/delivery between UE and server. The option may be transparent to 3GPP, and it can be left to implementation

**Q2: Do companies agree that these options can be used for RAN2 study? Please provide your comments in the comment column if any.**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
| OPPO | Yes | Just wonder whether CU/DU architecture should be considered also for Option1 even though RAN3 work may be involved. In our view, a note can be added for clarification:   * Option 1: Model transfer/delivery between UE and gNB. For this option, CP and UP solutions can be studied   Note: For Option1, both split and non-split gNB architecture may be considered. |
| Qualcomm | No | For options 1 – 3, we think the CP-based solution will not work. In our contribution paper [R2-2212659](https://www.3gpp.org/ftp/tsg_ran/WG2_RL2/TSGR2_120/Docs/R2-2212659.zip) we discussed it in detail. See further arguments below.  For option 4, although the model delivery is transparent to the 3GPP network. There will be associated procedures where gNB may be involved. For example, model identification, selection, and other LCM. Therefore, these aspects still need to be disused for option 4.  Now let us look into issues with CP-based solutions and the benefits of UP based solution over CP solutions as illustrated in the figure below.    From the figure above, we can observe that during mobility in poor channel conditions, different model delivery method has the following impacts:   * **AI/ML model is included in the configuration message**: it will make the RRCReconfiguration message significantly large. Many a time it may result in radio link failure, as the large configuration file may not be successfully delivered in poor channel condition in a timely fashion. * **AI/ML model is sent in another configuration message while the handover command is sent in the RRCReconfiugration message**: If the model is not transmitted successfully before the handover initialization, then AI/ML model needs to be transmitted again in the new cell. Note that upon handover control plane is terminated with the source gNB and initiated with the target gNB. Therefore, segments transmitted from the source gNB are considered lost, unless and otherwise, the target gNB retrieves the information about which segments have been successfully transmitted to the UE from the source cell. This will induce Xn signaling overhead. * **AI/ML model is configured in the RRCReconfiguration, and AI/ML model is delivered to the UE using the user plane**: it alleviates the probability of radio link failure and at the same time achieves seamless model delivery across gNBs. As the model delivery is between UE and the centralized server, then even in the case of handover failure or radio link failure, upon the connection establishment the model delivery can resume with the need for transmitting already received segments at the UE.   This clearly shows that model transfer using CP will result in frequent radio link failures and many a time complete retransmission of the models. Therefore, we think CP-based model transfer is not desirable for all options 1 – 3.  CP-based model transfer method has the following issues:   * Significantly high control plane overhead, as a large model size may need segmentation/transmission/acknowledgment. This consumes critical configuration time for model transfer/delivery. * Processing load during mobility at the gNB for delivering the model; associated with model segmentation and acknowledgment procedures. This may consume critical radio resources at the gNB. * Transmission of the configuration message containing the AI/ML model should not block other high-priority control messages. Therefore, any SRBs (e.g., SRB0, SRB1, and SRB3) carrying configuration messages should not be used for model transfer/delivery.   + In particular, during the mobility, the model should not be included in the RRCReconfiguration message as the channel condition may already be poor (in general, when RRCReconfiguration carrying handover command is sent channel may already be poor). We may want to avoid sending large configuration messages (containing AI/ML models) in such poor channel conditions. * An incomplete control plane model transfer has to be restarted upon mobility, as there are no current procedures to resume transmission across gNBs. * In option 1, gNB would have to store all the models for delivery as opposed to u-plane which supports centralized storage across many gNBs for the samemodel.   Furthermore, note that gNB may want to update models for more than one use case simultaneously, this will further create issues. The sum of the model sizes may be significantly large when models for multiple use cases needs to be delivered. |
| vivo | Yes with comments | For Option 1:  Currently, User Plane data can be exchanged between UE and gNB over DRB. However, the UP data is not terminated at gNB and will be further delivered to UPF. That means if the data needs to be terminated to gNB, the data must be rerouted back from UPF to gNB. This seems a complex data back and forth from gNB to UPF and UPF back to gNB.  Second, if the intention of UP solution in Option 1 is introducing a new User Plane date terminated at gNB This seems to break away from the current DRB data transmission design.  So we think UP solution in this option1 should be deprioritized.  For Option 2:  The LMF is also a CN entity. We suppose the intention is to distinguish LMF from other core network nodes as it’s for positioning only. Option 2 can be clarified as CN (except LMF).  For Option 3:  The UP Connection between LMF and UE has been studied and captured in TR 23700-71. Thus we think Option 3 shall also include the UP solution.    Besides, we can discuss the model transfer/delivery from NW to UE first. |
| NEC | See comments | In our view, each of the different options as listed by the moderator is based on an AI/ML architecture assumption behind. We suggest to discuss and sort out these assumptions before going to the discussion on the solution options.  Option1 may assume the AI/ML model training is done at gNB. Then for DL, the gNB transfers the trained model to the UE via air interface (CP or UP based).  Option2 may assume the AI/ML model training is done at CN. Then for DL, the CN transfers the trained model to the UE via NAS signalling or QoS flow based tunnel.  Option3 may assume the AI/ML model training is done at LMF. Then for DL, the LMF transfers the trained model to the UE via LPP. There may be other alternatives for this type of model transfer.  Option4 may assume the AI/ML model training is done at a specific server. Then for DL, the specific server transfers the trained model to the UE in OTT manner.  We suggest RAN2 to acknowledge the abovementioned AI/ML architecture assumptions before diving into the detailed solution discussion on the model transfer options, since this discussion can help to restrict the solution options. Otherwise, this email discussion seems automatically acknowledged all of the AI/ML architecture assumptions as listed. It would then be questionable if RAN2 should study the solutions for all of these AI/ML architecture assumptions.  For example, if the AI/ML model training is assumed only done at gNB, option 1 should be the focus of the study. And then we can simply evaluate the CP and UP based model transfer over air interface at the next step, which may help RAN2 to concentrate for this topic.  *PS: just provide a background, in RAN3 study on AI/ML, the assumption is based on that the model training is either at gNB or OAM entity.* |
| Apple | Yes with comments | For Option 1, we are not sure how UP solution work. In our understanding, UP solution is conflicted with "between UE and gNB" (i.e. terminated in gNB) because UP solution should be terminated in UPF (per current definition of UP in SA2). To able such UP solution work, SA2 spec is required to add a new establishment trigger type of UP tunnel (i.e. triggered by gNB). Since SA2 is not involved in this SI, we suggest to preclude UP solution in Option 1.  For Option 2, our understanding is same as vivo that it should be "CN (except LMF)".  For Option 3, we are confused by the so called "LCS-UP connection" provided by vivo (in TR 23700-71). We have two questions:  1) If "LCS-UP connection" means step 3 and 4 used to convey configuration between LMF and UE, we think it is still NAS signaling (i.e. CP solution)  2) If "LCS-UP connection" starts after Step 7, we are confused whether the UP connection is still between UE and UPF? Note that in Step 5, UPF is still involved. Then, if UPF is still involved in "LCS-UP connection", we are not sure what is its benefit over pure UP solution.  In addition, since this solution is just in SA2 TR, we think it is better to wait SA2 progress.  Finally, we are not convinced by Qualcomm's analysis on CP solution. We think more analysis with typical number as example is necessary, rather than just listing possible Cons of CP solution. At this stage, we think both CP and UP solutions should be studied.  Thus, we suggest below wording change:   * Option 1: Model transfer/delivery between UE and gNB. For this option, CP ~~and UP~~ solutions can be studied * Option 2: Model transfer/delivery between UE and CN (except LMF). For this option, CP and UP solutions can be studied * Option 3: Model transfer/delivery between UE and LMF. For this option, CP solution can be studied. Whether to study UP solution needs to wait SA2 input. * Option 4: Model transfer/delivery between UE and server. The option may be transparent to 3GPP, and it can be left to implementation |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

For this R18 Study Item, 3 main use cases are mentioned in the WID [1], i.e. AI/ML for CSI feedback enhancement, Beam management, and Positioning accuracy enhancement.

For each use case, it may correspond to one or more suitable architectural assumptions as listed above. For each architectural assumption, there may be some differences on solutions for applicable use cases. Table 2 is the email rapporteur’s understanding, based on the contributions at RAN2#120.

**Table 2: The relations between the architectural assumptions and applicable use cases**

|  |  |
| --- | --- |
| **Architectural assumptions** | **Applicable use cases** |
| Option 1 | CSI feedback enhancement  Beam management  Positioning accuracy enhancement |
| Option 2 | CSI feedback enhancement  Beam management  Positioning accuracy enhancement |
| Option 3 | Positioning accuracy enhancement |
| Option 4 | CSI feedback enhancement  Beam management  Positioning accuracy enhancement |

**Q3: Do companies agree on the relations in Table 2? Please provide your comments in the comment column if any.**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
| OPPO | Yes | For UP solution in Option1, this option is more challenging than CP solution in Option1 as current spec will not allow gNB alone to establish DRB resources for a specific UE without session establishment request from CN, but our understanding is that this question was set to confirm the options from very high level without touching any solution details, so we can share the pros/cons in Ph2 discussion.  Even though we also think it’s a little bit strange to keep AS AI/ML models at CN and use CP solution (e.g. NAS signalling) to transfer/delivery model between CN and UE, anyway model transfer/delivery method is a common topic not only aiming for RAN1-led three use cases, the future proof can also be considered if other high layer AI/ML use cases are introduced in the future, so let’s keep this option now.  As for the other Options, it’s more aligned with legacy spec design logic, nature to discuss further in Ph2. |
| Qualcomm |  | As shown above, all of the options UP-based solution are desired. Option 3 should include UP based solution |
| vivo | Yes, for  Option 3/4.  For Option 1/2, see comments. | For positioning, we think it’s straightforward to manage the model by LMF. Without “Positioning accuracy enhancement” use case, we are fine with Option 1 and Option 2. |
| NEC | No | See our comments for Q2. |
| Apple | Yes | We basically share the same view of rapporteur. The detailed analysis can be done in phase 2. |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## 2.2 Phase 2

### 2.2.1 Model transfer/delivery in Downlink and Uplink

*Note: According to model transfer/delivery defined by RAN1, the model transfer/delivery can be DL (NW to UE) or UL (UE to NW). Based on contributions at RAN2#120, it is observed that there were more interests in model transfer/delivery in DL than in UL, so it is suggested to collect companies’ views on the two directions.*

### 2.2.2 CP-based solutions

*Note: Based on the outcome of phase 1 discussion, this part is to collect companies’ views on principles, basic flows, and pros/cons for each possible option.*

### 2.2.3 UP-based solutions

*Note: Based on the outcome of phase 1 discussion, this part is to collect companies’ views on principles, basic flows, and pros/cons for each possible option.*

# 3 Conclusion

[To be added]

# 4 References

[1] RP-221348, Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface