**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 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# 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. |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

### 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. |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

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. |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## 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