**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 |
| Lenovo | Congchi Zhang | zhangcc16@lenovo.com |
| Xiaomi | Xing Yang | Yangxing1@xiaomi.com |
| Fujitsu | Xin Wang | wangxin@fujitsu.com |
| CATT | Erlin Zeng | erlin.zeng@catt.cn |
| China Unicom | Shuai Gao | gaos30@chinaunicom.cn |
| Huawei, HiSilicon | Jun Chen | jun.chen@huawei.com |
| Mediatek | Yuanyuan Zhang | Yuany.zhang@mediatek.com |
| CMCC | Jiayao Tan | tanjiayao@chinamobile.com |
| NTT DOCOMO | Souki Watanabe | souki.watanabe.gf@nttdocomo.com |
| ZTE | Fei Dong | Dong.fei@zte.com.cn |
| Nokia, Nokia Shanghai Bell | Sakira Hassan | sakira.hassan@nokia.com |
| LGE | Soo Kim | [soo.kim@lge.com](mailto:soo.kim@lge.com) |
| Spreadtrum | Xiaoyu Chen | xiaoyu.chen@unisoc.com |
| Ericsson | Felipe Arraño Scharager | felipe.arrano.scharager@ericsson.com |
| Intel | Ziyi Li | Ziyi.li@intel.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. |
| Lenovo | Both are fine if it’s only for the sake of discussion | We understand this is only for the sake of discussion in RAN2, as stated by the rapporteur.  We acknowledge the definition agreed in RAN1 and better we don’t introduce anything conflicting in RAN2. For the discussion purpose, we are fine with either way as long as RAN2 companies have the same understanding, i.e., “delivery” will become “transfer” if it’s sent via air interface to UE.  Eventually when we start drafting TR, we need to align the wording with RAN1 case by case following above definition. |
| Xiaomi | None | We understand the two terminologies are different and should not be mixed. Model delivery includes delivery between either two entities, which could be via air interface, e.g. between UE and gNB, or CN interface, e.g. between gNB and LMF. While, model transfer focuses on the deliver via air interface, i.e. between UE and other entities. In this email discussion, seems we only discuss the model delivery between UE and other entities. So, we can just use model transfer in this email discussion. |
| FUJITSU | Both are fine, preferable option 2. | At the beginning of the discussion of RAN1, the original terminology was only “model transfer”. Some companies add the term “delivery” for the sake of generality, especially enabled the possibility to transmit the model via entities beyond air interface such as OTT server. Therefore, we do not think there are huge differences between these two options.  If it is mandatory to choose one option, we prefer option 2, which shows coherence between the past and current discussions, and “model transfer” has not been officially deleted or banned from RAN1 discussion, the latest agreements still use this word so there is no reason to drop it at this stage. |
| CATT | Option2 | As has been pointed out by some companies, this is perhaps not a critical issue but we understand that this discussion is motivated by fact that in RAN2’s previous discussions there were cases where these two are somewhat misused. So, seems useful to align.  Generally speaking, model delivery is a term referring to delivery of an AI/ML model between entities, and model transfer refers to delivery mechanism that has air interface impact. RAN1’s definition is itself clear. And it should be clear that RAN2 do not need to spend a lot of effort on the mechanisms that are transparent to the air interface. |
| China Unicom | Option 2 | The term “model delivery” is more general than the term “model transfer”, but the latter is more accurate when we discuss model delivery over the air interface. So it’s better to use option 2 in this email discussion until RAN1 has made final decision. |
| Huawei, HiSilicon | Both are fine, slightly prefer option 2 | Firstly, we agree with other companies that Q1 is just about terminology issue, and we should avoid confusions.  Secondly, we think solutions for model transfer/delivery are open for now, and RAN2 can study them in this SI.  Option 2 is more generic so we also slightly prefer it. |
| Mediatek | Both are fine | Which terms to use depends on the scope of this email discussion. If we only intend to discuss model delivery over air interface, we can use the term model transfer; if we intend to discuss model delivery among any entity which may be beyond RAN2 and 3GPP, we can use the term model delivery. It’s not a critical issue if the term is only used for this email discussion and the scope o the email discussion is clear.  If we use ‘model transfer/delivery’ for discussion, it would be good to capture the reason somewhere why RAN2 use this term instead of model transfer or model delivery as defined by RAN1, in case that the term continues to be used in further discussion and liaisons to RAN1. |
| CMCC | Option 2 | Although model delivery can cover model transfer from the perspective of definition, we prefer option 2 during discussion phase since it reflects model delivery/transfer explicitly and is easier for understanding. |
| NTT DOCOMO | Option 2 | A distinction should be made according to the definition of RAN1.  For example, in considering the Q2 option in more detail, we will use the term "model delivery" because we will be discussing the difference in the functional part of the core side.  In addition, as can be seen from the following agreement in RAN1 for network-UE collaboration levels, RAN2 also notes that model transfer is not transparent to 3GPP signalling over the air interface.  RAN1 agreements  Take the following network-UE collaboration levels as one aspect for defining collaboration levels  • Level x: No collaboration  • Level y: Signaling-based collaboration without model transfer  • Level z: Signaling-based collaboration with model transfer  Working Assumption  • Define Level y-z boundary based on whether model delivery is transparent to 3gpp signalling over the air interface or not.  • Note: other procedures than model transfer/delivery are decoupled with collaboration level y-z  • Clarifying note: Level y includes cases without model delivery. |
| ZTE | Option 2 with comments | Terminology issue.  Option 1 seems not appropriate, model delivery just indicates the model exchange between different entities, and different combos of entities is not for sure, so we need use more precise terminology for indicating the different scenarios, for example:   * For the case that model exchange via air interface, the model transfer need to be dedicatedly used. * For the case that model exchange between two different entities other than model transfer (i.e. in most case, RAN3/SA2 shall be involved), we can use model delivery without air interface to indicate . |
| Nokia, Nokia Shanghai Bell | Option 2 | Option 2 considers both the definitions of model transfer and model delivery. |
| LGE | Option 2 | We have already assumed RAN2 reuses the terminology defined by RAN1 in the #119b-meeting. We prefer to use model delivery/transfer for the study until further input from RAN1 to avoid confusion between RAN1 and RAN2. |
| Spreadtrum | Option 2 | Option 2 is preferred. Because it reflects both the general cases where AIML model is exchanged between two entities and the specific cases where AIML model is exchanged over air interface.  And it is just a terminology issue, we can align with RAN1 conclusion in WI phase. |
| Ericsson | Option 2 | For now, Option 2 seems OK, as per the agreed use cases RAN2 needs to study both aspects, i.e., feasibility/ways of transferring an AIML model over the air interface, and feasibility/ways of delivering an AI/ML model from one NW entity (UE included) to another. |
| Intel | Option 2 | In our understanding, **model delivery** might cover 1. **model transfer** between UE and NW over air interface and 2. deliver model within network side (i.e. between NW nodes). From RAN2 point of view, the 2nd understanding of model delivery may not be our RAN2 scope to define signalling procedure (but it would help us to understand overall flow and procedure, see our response in Q2). Therefore, from air interface point of view, we think there’s no big difference between two terminologies, as we focus on how to deliver/transfer model from network to the UE if collaboration between network and UE is not transparent to 3GPP signaling and model delivery/transfer is required.  Also, as seen in RAN1 agreed collaboration level definition, model transfer and model delivery are used interchangeably. We don’t think there’s a need for RAN2 to waste time and debate on the difference.  Therefore, during the discussion, to avoid any ambiguity, we prefer to use option 2 “model delivery/transfer”, which can help us to discuss the overall procedure. |

### 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 | Yes, with comments | 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. Furthermore, we believe that there should be a unified solution for model delivery for all use cases for CN to UE. Also, there are existing solutions in SA2 for model storage, and other model management functions. Therefore, in our view, option 3 is not required  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.  In summary, we see the following two option as feasible solutions:   * Model transfer/delivery between UE and CN. For this option, UP solutions can be studied * Model transfer/delivery between UE and server. The option may be transparent to 3GPP, and it can be left to implementation. However, the associated procedures should be studied. |
| 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.  [Rapp] We see that some companies are not clear about UP solution for Option 1, and then they need to first understand how it works. At RAN2#120 meeting, some contributions mentioned the UP solution, e.g. R2-2211192, AIML Methods Discussion in General, OPPO (see Option5, Figure 5 and relevant text). The email rapporteur thinks that the principle of this UP solution is to use DRB for transferring AIML model(s) between UE, and the signalling procedures are not very clear for now (or there are some possibilities). So more companies’ views would be helpful.  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 |
| Lenovo | Yes with comments | First, we assume all options includes both UL and DL, since it’s under discussion now if UE shall transfer the model to NW as well.  For Option 1, how UP solution works is unclear to us neither, do we need to define some special DRB that terminates at gNB? Maybe it’s something worth discussion in phase 2.  For Option 2, agree with Vivo to exclude LMF to differentiate from Option 3, since LMF is part of CN.  For Option 3, we are also supportive to include the UP solution in the study scope. Even in legacy, SUPL can be used to transfer LPP messages as well, which is via UP. Besides, SA2 is going to support proper UP between LMF and UE in near future as Vivo commented. |
| Xiaomi | Comments | We think the one candidate solution is missed in option 1, which is the new AI layer between gNB and UE. We think this should be listed also as an option.  Furthermore, we wonder whether it possible that the model delivery is done between two entities without UE involvement. For example, the AI model is provided to gNB by other CN entities or server. This solution may be beneficial in gNB/LMF sided AIML model. In such solution, the AI model delivery between CN entities and gNB should also be considered. Although it may be out of RAN2 scope, the impact should be studied for the sake of whole solution. Maybe we can send LS to other groups if needed. |
| Fujitsu | Yes | For option 1, we agree with VIVO, the UP solution should be deprioritized at this stage.  For option 2, both UP and CP solutions can be considered. It is not so important to deliberately emphasize CN (except LMF) because we think the further study will be finally carried per use case, so when we talk about AI4Pos, we will naturally discuss both LMF and potential options for other CN entities, while for other use cases, LMF will not be considered.  For option 3, we also noticed that there are works to discuss how to deploy User Plane solutions between UE and LMF in recent specifications, so we suggest keeping this option, but the priority should be CP-based solution.  For option 4, it is OK. |
| CATT | Yes | Basically we see what the Rapp tries to do is to have a comprehensive list of the possible mechanism options, which may or may not have impact to RAN2. We are open to further discuss these options, e.g., to see which use cases they may be used for and to identify the specification impact for them if any.  Then we have the following comments on these options.  For Option 1, we also think that UP solution for Model transfer/delivery between UE and gNB seems not very clear. We need to first understand how it works. Perhaps we could first look into CP based mechanism for Option 1.  Regarding Option 2 and 3, we tend to agree some previous comments that LMF (Option 3) should be removed from other CN node(s), to avoid overlapping between Option 2 and 3.  For Option 4, it is transparent to 3GPP, and the model transfer/delivery can be left to implementation without any specification impact. So we do not expect many discussions for this one but it is OK to list it for the sake of completeness.  Additionally, we would like to point out that one more option may be added for the sake of completeness. As far as we know RAN1 discussions also include a possible way for “Positioning accuracy enhancement” use case, i.e., the “Case 3a: NG-RAN node assisted positioning with gNB-side model, AI/ML assisted positioning”. It is a gNB-side model solution, where the model may be delivered from LMF to gNB. So we may add the following to the list   * Option 5: Model transfer/delivery from LMF to gNB. For this option, CP and UP solutions can be studied. |
| China Unicom | Yes with comments | For option 1, both CP and UP solutions have different advantages in different cases, so the details can be discussed in the phase 2 before we decide to preclude any solutions, e.g., we need to further discuss the prerequisite that include the model size, model delivery direction (NW->UE/UE->NW), and how we can use the solutions in the commercial cases.  For option 2, we agree with vivo’s proposal to exclude LMF.  For option 3, support to include UP solution as SA2 has already integrated it in the TR.  For option 4, support. |
| Huawei, HiSilicon | Yes with comments | For option 1:   * As commented by OPPO, CU/DU architecture is suggested for the study, and we think it may be too early. For now, we think RAN2 can focus on network entities, and try to undertand the basic procedures for solutions * For UP solution, we share some companies’ views that how it works is unclear, and it bring some challenges to existing procedures, e.g. PDU Session/DRB establishment. In general, we see that option 1 UP solution may be complicated and bring lots of impacts to the standard, so we think the option may be de-prioritized   For option 2, we are ok with Apple’s suggestion, i.e. add “(except LMF)”.  For option 3, thanks to vivo for pointing out the latest SA2 progress on the UP solution. We do not observe any RAN2 or RAN3 impacts for the UP solution. In addition, due to the SA2 study, the UP solution is one of candidate solutions. In general, we think we can wait for more SA2 progress, and it seems not needed to list the UP solution for now.  In addition, we have more comments:   * We see that some companies provided detailed analysis (pros/cons) for some solutions, which are helpful and can be further discussed in Phase 2 * For NEC’s comments on the training part, our views are firstly the training entity is not restricted for now, and there are some possibilities, and secondly we think the discussions on model transfer/delivery solutions could be independent of discussing where the training entity is. For example, for option 1, the training entity may be gNB/CN/OAM/others, and then there may be some requirements on model delivery |
| Mediatek | Yes with comments | For option 1, although both CP and UP solutions are both possible, just as commented by vivo and Apple, it’s unclear how UP solution works in companies view. It’s possible that a new type of DRB can be defined and terminated in gNB. Considering the compatibility with current architecture, we prefer to start with CP solution first and consider UP solution later if time allows.  For option 2 and option 3, we prefer to keep option 2 as one generic option for model transfer between UE and CN and consider option 3 as a special use case of it. Because option 3 is not a generic assumption, which is purely applicable to AI/ML for positioning and very likely to be discussed in the use case specific aspect discussion.   * 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:~~ Consider the use case specific Model transfer/delivery between UE and LMF for positioning~~. For this option, CP solution can be studied~~ * Option ~~4~~ 3: Model transfer/delivery between UE and server. The option may be transparent to 3GPP, and it can be left to implementation |
| CMCC | Yes with comments | In general, we think both CP and UP solutions can be studied and evaluated at this early stage. And pros/cons for each possible option can be analyzed in phase 2. In addition, we assume that model training can be done at UE/gNB/OAM/CN or other entities, thus all options may include both UL and DL.  For option 1, UP solution is more challenging and may means complex data back or to introduce new special DRB which terminates at gNB as pointed by some companies, but we are open to discuss the solution.  For option 2, we also agree to exclude LMF from CN.  For option 3, since SA2 is studying the UP solution, we think it is better to wait for SA2 progress.  For other options, we think that model delivery between gNB and OAM or AMF can also be considered. For example, AI/ML model for CSI feedback or beam management is trained at OAM and then delivered to gNB, which is similar to the solution that model training at OAM and model inference at NG-RAN in RAN3 TR 37.817, or AI/ML model for positioning is trained at LMF and then delivered to gNB. We share similar view with Xiaomi and CATT to list these options for sake of completeness.  *- Option 5: Model transfer/delivery from OAM to gNB.*  *- Option 6: Model transfer/delivery from LMF to gNB.* |
| NTT DOCOMO | Yes with comments | For option 2/3  If a distinction between Option 2 and Option 3 is made at this stage, it is for use case considerations. However, we do not think the distinction between 2 and 3 is necessary for signaling from the RAN2 perspective.  For option 3  Since using CP/UP will be organized in phase2, we think that Option3 UP can be left as an option.  For option 4  Option 4 should be left as an option because it is necessary to consider how other LCM, e.g., model registration management, etc., will be different from the other options. |
| ZTE | Yes, if option 4 is excluded: | The model transfer issue is just identified by ourselves, RAN1 have no any guidance for RAN2 discussion so far. To evaluate each solution on the table, we need RAN1 to answer at least the following questions:   * how often the model transfer will be occurred? * what is the size of the model that needs to be transferred in most case? * What is the format of the model that needs to be transferred (i.e. open format or proprietary format, or both)   So, at this stage, option 1,2,3 on the table and other potential ones, if any, shall not be precluded.  However, we have some comments with option 4, to our understanding, the option 4 is not related to the model transfer, the connection between UE and their proprietary server is somewhat like normal data transmission, which is transparent to the NW, and also as rapporteur indicated, that is out of 3GPP scope, there is no need for RAN2 to spend time to discuss. |
| Nokia, Nokia Shanghai Bell | Yes with comments | In our view, all options should consider the identification of the model and the functionality to enable the LCM operations the model. Moreover, we suggest to analyze whether these options are viable solutions for the use cases (total 6 sub use cases) endorsed by RAN#98 plenary.  Option 1: Both CP and UP options can be studied, but the study should also explore the size of the models in question. We are also concern about the segmentation of critical messages due to the inclusion of large models in these solutions which requires further study.  Option 2: Yes. We agree with other companies that the LMF should be explicitly excluded from CN in this option since it is treated separately in option 3.  Option 3: We should consider CP and UP solutions.  Option 4: Yes. If model transfer/delivery between a UE and server strictly applies to the download or upload, then the download or upload could be transparent to 3GPP. Moreover, we suggest to consider the requirements for collaboration level ‘x’ and ‘y’ for this option.  We appreciate the comments from other companies which will be useful in studying Phase 2. |
| LGE | Yes with comments | For Option 1, as some companies mentioned above, UP solution is not suitable considering the data exchange between gNB and UPF. So, we think UP solution should be deprioritized in Option 1.  For Option 2, we believe both CP and UP solutions should be studied to consider all of model transfer/delivery scenarios. |
| Spreadtrum | Yes with comments | In this early and open stage, we think any candidate solution that RAN2 involved should not be precluded. And we want to emphasise that the model transfer/delivery options may different for different use cases, like for CSI and positioning. And the model transfer/delivery options maybe also different when considering model size and latency restriction etc.  For Option 1, as pointed by some companies, the UP solution may be complex to define the special DRB terminated at gNB. We slightly prefer to deprioritize the UP solution, but in this early stage, both can be further studied.  For Option 2, we support to exclude LMF from CN for the purpose of distinguishing Option 2 and Option 3.  For Option 3, thanks for vivo’s clarification, for UP solution we can align with SA2 progress.  For Option 4, it is probably the common method for model transfer/delivery. Because it can be used for the cases including AI4NW and NW4AI (application oriented). As rapporteur pointed, Option 4 should be left to implementation. |
| Ericsson | Yes, with comments | Acknowledging that there is no “clear picture” in RAN2 (e.g., missing requirements/details) which could further complexify the discussion’s scope and the design of solutions … For general purposes RAN2 could start by focusing on Option 1 and limit the analysis to aspects within our WG reach, i.e., to RAN2-specific entities, procedures, and protocols. This, since neither RAN3 nor CT/SA groups have TUs allocated to this SID.  For Option 2 and, especially Option 4, RAN2 should not spend time discussing, agreeing or even capturing implementation-specific solutions that are transparent to 3GPP signalling or RAN2 specification.  Option 3 can directly be addressed in the positioning (sub)use case discussion. |
| Intel | Yes with comment | In our understanding, methods supporting model delivery/transfer between UE and network should be applicable for both UL and DL. Therefore, the discussion is suggested to cover both UL and DL.  For the options listed above, we share the similar view with rapp that it could be the starting point and further study and evaluation could be done during SI phase.  However, we share the similar view with NEC and CMCC that the approach for model delivery/transfer highly rely on where model training taken place.  Assuming the options listed above showing the start (where the model is generated) and destination (where the model is used) of model transfer/delivery, if model is trained at CN, for example, there’s less motivation to consider Option 1, where model is delivered/transferred between gNB and UE. Considering it is also RAN2 responsibility to study the allocation of functionalities to entities, we think it’s good to jointly consider model delivery/transfer solutions with functionality mapping.  Here are our comments regarding to different options:  For Option 1: as commented by other companies, using UP solution between gNB/UE is complicated, as gNB cannot establish DRB by itself (i.e. DRB is terminated at CN). Therefore, how to support UP Solution in option 1 is not clear and require further study.  For Option 2: agree with Vivo and Apple’s comment, this could be further updated to CN (except LMF)  For Option 3: Though SA2 is discussing UP Solution between LMF and UE, from the figure, it seems the model delivery/transfer is transparent to NG-RAN, i.e. no RAN impact. We think it’s ok to only consider CP solution at this stage. Once SA2 has further progress and identify any RAN impact on UP solution, we can further study correspondingly.  For Option 4: we are wondering whether the serve in Option 4 also include OAM? It is noted that model training at OAM and gNB are considered in RAN3 WI “AI/ML at NG-RAN”. Hence, if OAM is not included, we think another option can be considered:  **Option 5: model transfer/delivery between UE and OAM. For this option, CP solution can be studied.**  In the end, for the completeness of the solution, we think it would be helpful to consider how the model is transferred/delivered within the network. Though RAN2 may not be the responsible WG, we could capture the assumptions during SI and check with other WGs. |

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. |
| Lenovo | Yes |  |
| Xiaomi | Yes |  |
| Fujitsu | Yes | It is quite straightforward to have this initial classification, CSI and BM may not need to transfer the models to CN or store them at CN, however, the lifecycle management could be a generic framework for all use cases, so we suggest to keep this classification at this stage. |
| CATT | Yes, for  Option 1/3/4.  For Option 2, see comments. | For Option 1 and 4 we generally agree all the identified sub use cases may use these model delivery/transfer mechanisms. Although we feel that based on what RAN1 had discussed, using Option 1 for ‘Positioning accuracy enhancement’ is a bit strange solution, e.g., does this mean UE needs to report its related measurements to gNB, which is different from current positioning mechanisms? But we are open to further discuss.  For Option 3, we also agree that currently it seems specifically targeted for the use case of ‘positioning accuracy enhancement’.  For Option 2 we are not sure which use cases should be listed for it. Technically speaking the functionalities of ‘CSI feedback enhancement’ and ‘Beam management’ only involve UE and gNB. It seems naturally to focus these use cases on Option 1. Likewise, it may be further discussed whether for ‘Positioning accuracy enhancement’ Option 1 and/or Option 3 are not sufficient. That being said, we are OK to further discuss Option 2 considering that this study item should be future proof and maybe in the future we will see suitable use cases for it.  Additionally, as said in the previous question we may add Option 5 which may be useful for ‘positioning accuracy enhancement’ (at least for the sake of completeness at this early stage of studies). |
| China Unicom | Yes | The relation table are fine for us, and we also agree to leave the detailed discussion in phase 2. |
| Huawei, HiSilicon | Yes | For CSI feedback and BM use cases, we are ok with the above analysis in Table 2.  For Positioning use cases, there are some discussions on model transfer/delivery options in Q2, which may impact the text in Table 2. |
| Mediatek | Yes with comment | Option 3 is a special case of option 2, which is necessarily listed as one option.  E.g., Option 2: CSI feedback enhancement, Beam management, Positioning accuracy enhancement (with LMF) |
| CMCC | Yes | In addition, as we commented in Q2, if option 5 and option 6 are added, option 5 can be applied for CSI feedback and beam management use cases, and option 6 can be applied for positioning use case. |
| NTT DOCOMO | Yes |  |
| ZTE | Yes, with comments | The row with option 4 can be removed from the table as we comment in the second question.  We would like to note that, this only can be assumptions in RAN2, we cannot achieve any conclusion regarding the use cases at this stage without any information about use cases from RAN1,  for the second stage, we think it is better for RAN2 to firstly identify the CONs and PROs for each option in a general way. |
| Nokia, Nokia Shanghai Bell | No, please see comments. | Option 1: This option may be insufficient for positioning use case.  Option 2: This option may be insufficient for positioning use case. We should separate LMF from the CN.  Option 3: Yes.  Option 4: For models coming from a server, it is unclear what level of collaboration is possible, which could make joint training of a CSI feedback enhancement model impossible, for example. For beam management and positioning accuracy enhancement, this is still feasible for a UE-side model to essentially augment legacy methods in a transparent manner. In addition to this, with Option 4, we need to consider the validation, authentication, and security. |
| LGE | Yes |  |
| Spreadtrum | Yes | The relations in Table is fine to us. Agree to have a further study in phase2. |
| Ericsson | Yes | For completeness, please refer to our answer to Q2. |
| Intel | See comment | For positioning use case, LMF is typically used for positioning in legacy without AI/ML. Therefore, we think for such use case, the model delivery/transfer method can be considered separately from the other two use cases, i.e. option 1/2 is not considered for positioning accuracy enhancement.  For the other two use cases, though option 1/2/4 may be applicable for CSI and beam management use cases, we need to discuss and identify where the model is trained/generated, which is the baseline to further discuss how to transfer it over the air interface.  Besides, as we commented in Q2, option 5 “model transfer/delivery between UE and OAM” could also be considered by all three use cases. |

## 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