**F3GPP TSG RAN2 #121 R2-23xxxxx**

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

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

**Source: Huawei (email rapporteur)d**

**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 |
| Samsung | Chadi Khirallah | c.khirallah@samsung.com |
| Futurewei | Chunhui (Allan) Zhu | chunhui.zhu@futurewei.com |
| Interdigital | Oumer Teyeb | oumer.teyeb@interdigital.com |
| Dell Technologies | Ali Esswie | ali.esswie@dell.com |
| Futurewei | Chunhui (Allan) Zhu | chunhui.zhu@futurewei.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** |
| Dell Technologies | Option 2 | Option 2 is more generic and includes AI/ML model exchanges over radio and non-radio interafces, and also avoids any potential misalignment with RAN1. It can be therefore understood that AI/ML model delivery/transfer referes to model transfer if it s over the radio interface.  We suggest RAN2 to indicate such reference with the termeonology agreement. |
| 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. |
| Samsung | Both are ok | In our understanding, RAN1 defined both terminologies “model delivery” and “model transfer”.  However, we agree that Option 2 “model transfer/delivery” is a more generic/unified term. |
| Futurewei | Option 2 | The problem with Option 1 is that it uses only the term “model deliver” but it tries to cover both “model transfer” and “model delivery”. This will cause unnecessary confusions. |
| Interdigital | Slight preference to Option 2 | Both options are OK, but option 2 will avoid further discussions on model delivery vs model transfer. |

**Summary:**

Most of companies prefer option 2.

For the solution discussions in this email discussion, some companies think RAN2 can discuss model transfer/delivery solutions between UE and other entities first.

**Proposal 1: Use the wording “model transfer/delivery” for the RAN2 study.**

### 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. |
| Samsung | Yes with comments | Overall, we are open to study specification impacts of different options with possible modifications/clarification suggested above (e.g. by vivo and apple):  For option 1, need to clarify how to use UP solution for model transfer/delivery.  For option 2, no strong view, we could exclude LMF since it is mentioned in Option 3.  For option 3, whether to study UP solution could wait for further progress /input from SA2.  For option 4, ok. |
| Futurewei | Yes, with comments | For Option 1, although, as vivo and some other companies pointed out, the UP solution involves more standards impact and hence more complicated, the CP solution has its own issues. For example, we are not sure whether it can easily handle large-size models, e.g., tens of MBs. We also share the concerns expressed by QC on CP solution during mobility events. Based on this understanding, we believe both solutions should be studied for R18. We can decide our next move based on the result of the study.  We agreed with vivo that LMF is part of CN. However, regarding whether to exclude LMF from Option 2 and make LMF an independent option or to make LMF (Option 3) a special case of CN (Option 2), the decision should be made by checking whether they share the same transfer/delivery mechanism.   * If they share the same mechanism (e.g., all use NAS over RRC) and just the use case being unique for LMF (for positioning only), then we think it would be OK to make Option 3 a special case of Option 2. * Otherwise, it is better making LMF an independent option (i.e., keeping it as Option 3).   We are fine with Option 4. |
| Interdigital | Yes | We are fine with all options as baselines for starting the discussion. As some companies have pointed out, there maybe some issues with UP and/or CP based solutions for some of the options. However, we think it is too early to down select UP or CP for each option. Some level of down selection or (de-)prioritization could be attempted during phase 2. |

**Summary:**

Option 1:

* For UP solution, how it works is not clear to some companies. vivo, Apple, Lenovo, Fujitsu, CATT, Huawei, MediaTek, LGE (8) think the solution can be deprioritized because it is unclear and it may be complex. The email rapporteur suggests to postpone the UP solution, due to unclarity of the solution
* Xiaomi proposes a new solution, i.e. the new AI layer between gNB and UE. OPPO wonders whether CU/DU architecture should be considered. The email rapporteur thinks this email discussion can focus on solutions with more interests from companies, and others can be postponed

Option 2:

* The LMF is also a CN entity, and it will be good to distinguish LMF from other CN nodes. Some companies are supportive with the change “CN (except LMF)”, and the distinction is made just for use case considerations

Option 3:

* vivo pointed out that a UP solution has been studied and captured in TR 23700-71 (SA2), and then the UP solution can be studied. Lenovo, China Unicom, NTT DOCOMO, Nokia (4) have the same view. However, some companies (Apple, Huawei, CMCC, Intel, Samsung) have a different view, and they think the solution is just from SA2 TR, so we can wait SA2 progress/input and we do not study it for now. The email rapporteur suggests that whether to study UP solution needs to wait for SA2 progress/input

Option 4:

* Most of companies are fine with the current text
* Qualcomm pointed out that the associated procedures may impact RAN2, e.g. model identification, selection, and other LCM. Therefore, these aspects still need to be disused for option 4. NTT DOCOMO have the same view. The email rapporteur thinks that this email discussion is mainly for model transfer/delivery, and other LCM aspects may be discussed in other places later
* ZTE pointed out that the option 4 is not related to the model transfer, and thus there is no need for RAN2 to spend time to discuss it

Other options:

* CATT pointed out that for positioning use case 3a, the gNB-side model solution may need model delivery, e.g. from LMF to gNB. CMCC pointed out that model delivery between gNB and OAM/CN can be also considered
* For discussions in Q1, it is observed that some companies prefer to discuss model transfer/delivery solutions between UE and other entities in this email discussion. The email rapporteur thinks that model delivery between network entities is not precluded but can be discussed later, so the email scope could be more focused
* For Option 5 proposed by Intel, it is the email rapporteur’s understanding that the model transfer/delivery between UE and OAM may be done via two ways. The 1st way is that if OAM is considered as a server, then it is similar to option 4. The 2nd way is that OAM can send AI/ML model(s) to some network entities (e.g. gNB/CN), and then the later ones send AI/ML model(s) to UE. For the 2nd way, the impacts to air interface would be similar to option 1 and 2. As commented by the email rapporteur above, the model delivery between network entities is not precluded but can be discussed later
  + **Option 5: model transfer/delivery between UE and OAM. For this option, CP solution can be studied.**

Others:

* Some companies provided detailed analysis (e.g. pros/cons) for solutions, which are helpful but the email rapporteur thinks that we can discuss the details in phase 2

Based on the analysis above and more comments from companies, the following summary proposals are made:

**Proposal 2: Agree to discuss the following solutions in phase 2:**

* **Option 1: Model transfer/delivery between UE and gNB via CP and UP solutions**
* **Option 2: Model transfer/delivery between UE and CN (except LMF) via CP and UP solutions**
* **Option 3: Model transfer/delivery between UE and LMF via CP and UP solutions**
* **Option 4: Model transfer/delivery between UE and server**

**Proposal 3: The discussion on model delivery between network entities is postponed.**

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. |
| Samsung | Yes with comment | For Options 1-3, ok to study / discuss in detail, for example, whether there is a need for model delivery in all use cases listed in the table.  In addition, for Option 4 (i.e. transparent to 3GPP), we are not sure whether RAN2 should/need to study any proprietary model delivery. |
| Futurewei | Yes, with comments | Agree in general.  Comment: the table may need to be adjusted accordingly based on the results of discussions of Q2 (whether Option 2 and Option 3 may be merged). |
| Interdigital | Yes | Agree. We can start the detailed analysis of option to use-case mapping during phase 2. |

**Summary:**

Most of companies are fine with Table 2, while there are some comments/suggestions.

* For Option 1 and Option 2, some companies (vivo, Nokia, Intel) think that both options should not be considered for Positioning accuracy enhancement, because LMF is typically used for positioning in legacy without AI/ML
* The email rapporteur thinks that both options can focus on CSI feedback enhancement and Beam management, and there are no specific considerations for Positioning accuracy enhancement. So Table 2 is updated by adding two notes

**Proposal 4: Agree on Table 2a for the RAN2 study and it can be used for further discussions.**

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

|  |  |
| --- | --- |
| **Architectural assumptions** | **Applicable use cases** |
| Option 1 | CSI feedback enhancement  Beam management  Note: No specific considerations for Positioning accuracy enhancement for this option. |
| Option 2 | CSI feedback enhancement  Beam management  Note: No specific considerations for Positioning accuracy enhancement for this option. |
| Option 3 | Positioning accuracy enhancement |
| Option 4 | CSI feedback enhancement  Beam management  Positioning accuracy enhancement |

## 2.2 Phase 2

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

According to model transfer/delivery defined by RAN1, the model transfer/delivery can be in Downlink (from network to UE) or in Uplink (from UE to network). The email rapporteur understands that for model transfer/delivery in Downlink, it is applicable for UE-sided AI/ML model, and for UE part model of two-sided AI/ML model; for model transfer/delivery in Uplink, it is applicable for network-sided AI/ML model, and for network part model of two-sided AI/ML model.

The email rapporteur observes that the workload for studying both directions will be much more than for studying one direction, so RAN2 may focus on one direction and discuss the other direction later. It is suggested to collect companies’ views on the two directions.

For RAN2#121, regarding the discussion on model transfer/delivery in Downlink and Uplink, two options are provided and can be discussed:

- Option 1: we start with discussing model transfer/delivery in Downlink first, and then we may discuss model transfer/delivery in Uplink later (based on companies’ contributions/preferences)

- Option 2: we discuss model transfer/delivery in both Downlink and Uplink

Note: In the following sections (e.g. 2.2.2, 2.2.3. 2.2.4), the email rapporteur suggests to mainly discuss model transfer/delivery in Downlink, and for Uplink, the interested companies can also provide inputs if any.

**Q4: Regarding model transfer/delivery in Downlink and Uplink, which option is preferred?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Option 1/2** | **Comments** |
| vivo | Option 1 | The network can train a reliable model for UE based on the collected data from multiple UEs. On the contrary, one single UE has a limited dataset or capability to train a trustworthy model for the network.  Besides, when multiple UEs transfer different models (e.g., different model structures and/or parameters) to the network, it will be a big challenge for the network to handle/maintain the models.  Therefore, the downlink model transfer shall be prioritized. |
| Xiaomi | Option 2 | As rapp points out, RAN1 considers AI model delivery in both directions. It’s better to consider both directions in the beginning. If only downlink is considered, we may end up with a solution doesn’t support uplink. Therefore, we think the solution which can support both directions should be prioritized. |
| Qualcomm | Option 1 | The requirement for model transfer/delivery in uplink is not clear, yet. Therefore, we believe that RAN2 should focus first on studying model transfer/delivery in Downlink. |
| LGE | Option 1 | We think that the UE can perform model inference on the model trained by the UE. However, considering the capability of the UE and the amount of information it has, the probability that the network performs inference on the corresponding model trained by the UE seems low.  Therefore, we can start with discussing model transfer/delivery in Downlink first. |
| Apple | Slightly prefer Option 1 | We agree with Rapporteur that we can first study downlink considering workload. However, we think it is necessary to clarify that it doesn't mean RAN2 think downlink model transfer (mapped to UE-sided model) is prioritized over uplink model transfer (mapped NW-sided model). Whether UE-sided model is prioritized should be decided by RAN1, and RAN2 agreement should not be misused to influence RAN1 discussion.  In addition, we disagree with some above comments on challgenges of uplink model transfer. We think the transmission efficiency issue (e.g. under multiple UEs) is within RAN1 scope, and RAN2 just need to focus on impacts of signaling. In our understanding, the main signaling difference between dowlink transfer and uplink transfer is that UE has less memory size than gNB, thus max number of signaling segements may be different between DL and UL. We can discuss uplink transfer issue after donwlink transfer signaling is clear. |
| OPPO | Option1 | Model transfer/delivery in Downlink is more typical than that in uplink even if most spec impact may be common in the end, at this early stage, we think it may be efficient to focus on Downlink first, if the analysis outcome can be easily port to uplink, we can do that on top of downlink. |
| Dell Technologies | Option 1 | We suggest to start the study of AI/ML transfer/delivery in the Downlink direction first, as it is the more common AI/ML framework. Based on the determined standardization impact of this study phase, we can move to the uplink direction, where a similar spec impact may be expected. |
| Huawei, HiSilicon | Option 1 | For now, it is not needed to exclude some directions, and the key question is how to discuss the directions. Due to limited TU in RAN2, we think it is important to focus on parts with more companies’ interests, so we prefer option 1.  For option 2, we think there are some commonalities between solutions for model transfer/delivery in DL and UL, so once there are sufficient discussions on DL, we can check details in UL. |
| NEC | Option 1 | We believe the model transfer from the network to the UE i.e. downlink based transmission is the mainstream use case. The requirement for uplink based approach (as a symtric handling of downlink) can be discussed when downlink based transmission is clear. |
| Lenovo | Option 1 | We also believe could be easier to start with DL first. The same principle could be applicable to UL as well, e.g., w.r.t CP/UP. On the other hand, that should not mean RAN2 can only discuss UL after DL discussion is completed. Once the use of UL model transfer becomes more clear (especially from RAN1 perspective), RAN2 can start the relevant discussion as well. |
| Spreadtrum | Option 1 | We agree to study model transfer/delivery in DL firstly, considering the gap between NW and UE in storage and computing, DL model transfer/delivery should be the mostly used scences in current stage .  The solution of Model transfer/delivery in UL may be similar with the solution in DL and can be discussed later. |
| CMCC | Option 1 | We think that it may be efficient to discuss model transfer/delivery in Downlink first considering the workload. As pointed by rapp, both Downlink and Uplink are in the RAN1’s scope, so we think RAN2 should clarify that it doesn’t mean RAN2 deprioritize/preclude Uplink model transfer. |
| Nokia, Nokia Shanghai Bell | Option 1 | The first option seems more preferrable due to the following reasons assuming that NW is hosting the model for training and UE is used for inference:  - UE can only train on local data collected by itself.  - NW has more information about multiple UEs for training if NW is used as training host.  - Less authentication and security validation are needed in the NW side.  - Most use cases in Rel-18 can support this option  - NW needs to manage fewer of models per use case  - Generalization is possible  - NW does not need to consider any specific configuration tailored to specific UE vendor.  - Model update is easier from NW perspective |
| ZTE | Option 1 | No matter for DL model transfer or UL model transfer, what RAN2 shall do is to find an available tunnel for model transfer between UE and CN, UE and LMF, UE and gNB. It is okay for us to start the study from the DL model transfer. If the tunnel for model tranfer is determined, to our understanding, it is not a hard work to apply the tunnel to the UL model transfer. |
| Ericsson | Option 1 | As mentioned by other companies, we have no clear understanding of how model transfer/delivery in the UL would work. It would of course be beneficial to have further details of what would be transferred/delivered and the purpose of such information from UE to NW.  On the other hand, we agree with others; the DL case seems more “common” from a RAN2 standpoint. Hence, it appears reasonable for RAN2 to focus on analysing the feasibility and the real need of having specified solutions to transfer/deliver a model in the DL first. |
| Fujitsu | Option 1 | We think both the DL and UP model transfer/delivery need to be studied at the end, but from the workload point of view, we prefer to select one of the direction for pilot study, and DL NW->UE will be a decent starting point due to:   * For some use cases such as positioning, only NWcan collect all the related data and assistance information (e.g., LMF collect information from multiple gNBs and UEs) to facilitate the model transfer. * It is natural to have typical applications for DL model transfer, e.g., model re-training/fine-tuning in NW and send it to UE for updating. * NW knows the overall situation for multiple models, it can handle the configuration and management among requests from more than one models.   Anyway, we agree with some comments above that RAN2’s focus should be the signaling and procedure to enable the model transfer/delivery, therefore UL study can reuse the results and conclusions once the DL study is sufficient. |
| CATT | Option 1 | We agree to start the discussing from Downlink model transfer/delivery first. And discussion on Uplink direction case can wait for more progress in Downlink direction, since they may share many aspects in common. |
| Futurewei | Option 1 | Agree with most companies to start with DL first. Given the limited TU assigned to this study, we need to be focused. |
| China Unicom | Option 1 | It’s common that the NW has more advantages in the storage, computing, power consumption resources when compared with the UE. So when we agree the rapp’s suggestion to start from the downlink-solution, which may be more likely to be commercially used in the network. |
| Mediatek | Option 2 | The intention of the email discussion is to collect cons/pros for different model transfer/delivery ways instead of justify which direction of model transfer/delivery is more pratical. The analysis and conclusion should not be much different for model download or model upload.  From procedure and signaling point of view, model download and model upload should share as much commonality as possible. Just as mentioned by several companies above, it is expected that model upload is similar as model download. Are there any paritular aspects identified for model upload which require different consideration from model download?  We are OK to discuss model download just as an example. But the conclusion should be considered as applicable to model upload unless the exceptional case is mentiond. |
| Samsung | Option 1 | We support breaking down the workload of the study item into two stages:   * First stage: model transfer/delivery in the downlink direction, then * Second stage: model transfer/delivery in the uplink direction.   We also think that RAN2 could consider signalling impacts in the case of model transfer/delivery from/to multiple UEs to/from the network. For example, it could be a case of transfer/delivery of the same model or different models (or models’ parameters) to/from multiple UEs.  In our view, generally at this stage, RAN2 should deprioritize or postpone the discussion on model transfer/delivery and wait for RAN1 conclusion on this issue. For example, whether the model transfer/delivery is really needed for each use case, or whether there is a need for model transfer. |

### 2.2.2 CP-based solutions

Based on the outcome of phase 1 discussion, this part is to collect companies’ views on principles, basic flows, pros/cons, and others for each possible CP-based solution. RAN2 impacts and cross-WG impacts can also be discussed. The following solutions have been identified for the RAN2 study due to phase 1 summary:

* Option 1 – CP solution that gNB can transfer/deliver AI/ML model(s) to UE via RRC signalling.
* Option 2 – CP solution that CN (except LMF) can transfer/deliver AI/ML model(s) to UE via NAS signalling.
* Option 3 – CP solution that LMF can transfer/deliver AI/ML model(s) to UE via LPP signalling.

#### 2.2.2.1 Option 1 – CP solution

For this CP solution, the principle is that gNB can transfer/deliver AI/ML model(s) to UE via RRC signalling. As discussed in phase 1, the applicable use cases are CSI feedback enhancement and Beam management, and there are no specific considerations for Positioning accuracy enhancement.

The basic flow for this CP solution is shown in figure 1 below.



Figure 1: Basic flow for Option 1 – CP solution

**Q5: Regarding Option 1 – CP solution, do companies agree with the principle and the basic flow described above?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
| vivo | Yes with comments | Figure 1 can be a baseline and the following procedures can be further discussed:   * The model transfer/delivery can be a procedure with a response. The response can be used for the UE to indicate the model is ready for inference/activation when UE has re-segmented/compiled the model. * The UE may request model transfer/delivery beforehand based on its capability and preference. |
| Xiaomi | Yes | RRCReconfiguration or new RRC message can be used for downlink.  There seems to be no appropriate existing RRC message and procedure for uplink. New RRC message may be introduced for uplink. |
| Qualcomm | Yes | See further comments in next question |
| LGE | Yes |  |
| Apple | Yes | We think whether new RRC message is needed and signaling details are stage 3 issue. |
| OPPO | Yes with comments | We think AI model is more than a new type of traditional configuration which is usually controlled by UE capability procedure, instead, AI model is more like a new kind of service, we think UE may need to get the network authorization on model transfer/delivery before transferring/delivering AI model to UE side; otherwise, operator may lose interest to training AI models itself or introduce AI models from third party for model transfer/delivery considering the balance between potential gains (e.g. system performance and new commercial growth point) and efforts. More addition, we think both network triggered and UE triggered model transfer/delivery should be considered, only network triggered model transfer/delivery is reflected into the Figure 1, better to clarify this also. |
| Dell Technologies | Yes | Where new RRC signaling messages can be defined for both downlink and uplink AI/ML model delivery/transfer. However, this can be left to stage-3 phase. |
| Huawei, HiSilicon | Yes |  |
| NEC | Yes | The figure can serve as the stage 2 baseline concept |
| Lenovo | Yes | Details can be discussed later. We also understand that does not mean the AI model is trained by gNB. It is also possible that the AI model is trained by OAM/CN and delivered to gNB. |
| Spreadtrum | Yes | The above figure 1 can be used as baseline for Option 1 – CP solution.  Other signalling enhancement like pre-signalling of authorization, capability indication etc, subsequent response signalling, and signalling form (via new or legacy RRC signalling) can be further discussed. |
| CMCC | Yes | We think figure 1 can be a baseline and whether a new RRC message is needed can be discussed in stage 3. |
| Nokia, Nokia Shanghai Bell | Yes with minor comment | As a baseline, this signalling flow is Ok and the details of the RRC signalling can be studied further. The main focus is to study the limitation of RRC signalling for model transfer/delivery at least in the downlink direction. |
| ZTE | Yes, with comments. | For this general procedure, we think we need to discussion the following issues:  1: What RRC message shall be used for model transfer, a new DL RRC message, or an existed DL RRC message.  2: Whether a new SBR shall be introduced for the model transfer?  3: What is the model transfer request-response procedure between UE and gNB. |
| Ericsson | Yes | Check our input to Q6. |
| Fujitsu | Yes | The basic flow is quite straightforward as the core of the CP solution so we have no reason to object it. And we think the first priority is to design the full mechanism around the basic flow, e.g., the request/feedback/error handle flows, for the details of SRB/RRC messages, it can be considered later. |
| CATT | Yes | And then it is necessary to consider how to convery the large size model, e.g. the segmentation of the downlink or uplink model. |
| Futurewei | Yes | This can be the baseline for Option 1. |
| China Unicom | Yes |  |
| Mediatek | Yes | It’s stage-3 issue on whether to reuse RRC reconfiguration message or define new RRC message. |
| Samsung | Yes | We have a similar view to VIVO and OPPO that a UE may request model transfer/delivery from the network that can decide which model(s) to be delivered to the UE. For example, the network decision may be based on UE capabilities, use case, etc. |

For this CP solution, it is suggested to collect pros/cons and other comments.

**Q6: Please provide your comments on Option 1 – CP solution in the table below, such as pros/cons, impacts due to model size/latency, use case specific analysis/comments.**

|  |  |
| --- | --- |
| **Company** | **Comments** |
| vivo | **Pros:**   * No multi-vendor interoperability issue if the model format/description is clarified during model transfer. That is, no offline coordination is needed between different vendors. * The gNB can update/transfer suitable models to UE with limited latency, especially during the handover, which will improve the overall user experience. * If the model is visible to the RRC layer, delta configuration can be used to reduce the signaling overhead.   **Cons:**   * [The segmentation number needs to be extended if the model size exceeds 40000Bytes.]   Note: If a separate AI layer on top of RRC is introduced, which can handle the segmentation/re-segmentation of the model, this con does not exist.   * [A new SRB type may be introduced to reduce the impact on the signaling transmission on other SRBs.]   Note: if SRB4 for QoE is reused, this con does not exist. |
| Xiaomi | Pros:   * Limited spec impact to support downlink model transfer. * Less latency compared with the LMF/CN solutions.   Cons:   * May be difficult to convey large size AI model by RRC message. * Uplink model transfer may require new RRC message and procedure * If the *RRCReconfiguration* message is used, AI model transfer may be impact by other radio configuration in the same message. For example, the compliance check failure of other radio configuration may result in AI model transfer failure in the same *RRCReconfiguration* message. |
| Qualcomm | We request companies consider the following when considering the solution for model delivery/transfer,   * Individual model size, i.e., model of an individual use case. * Cumulative model size, i.e., total model size of different use cases that may be required to be delivered or transferred to the UE, simultaneously. * Model delivery/transfer primarily will be required during the handover when channel conditions may be already significantly poor. * Model duplication may be required at every gNB. * F1 overhead in the split gNB architecture.   Now, if we consider two different methods, (i) Model is included in the RRC(Re-)Configuration message, and (ii) model and RRC(Re-)Configuration message are sent separately over the control plane. We have the following issues as shown in the figure below:  Diagram  Description automatically generated  As can be seen from the figure above, the following issues can be observed,   * **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.   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, a CP-based solution has the following issues,  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.   Significant F1 overhead in the split gNB architecture. |
| LGE | Pros :   * gNB can consider a ML model configured for a UE as UE contrext, which allows existing UE context transfer from source to target to be applicable for mobility   Cons:   * gNB is in charge of configuring a propoer ML model, which in turn requires the gNB to have a complete understanding of the ML model to use. This then requires massive update of existing gNBs to support ML functionalities. * Transfer of a big model causes several issues   + Segmented delivery of the RRC message is needed to carry a big ML model. In case many sgements are needed, it incurs significant transport latency.   + To avoid head-of-blocking of time-critical SRBs, SRB4 or a new SRB of lower priority should be used to send the RRC message carrying the ML model. * RAN2 spec impact is significant, and RRC cannot catch up the latest ML model pool applicable in the field/ML community.   + RAN2 needs to standardize ML models in RRC. For any model to use, its network model and model parameters shall be specified. This would introduce a strong but absolutely unnecessary dependency between RRC and ML syntax.   + The pool of applicbale ML models in the field is rapidly evolving and expanding, whereas the change is RRC spec is quite slow and strictly controlled. This means that RRC can never catch up the progress of ML models applicable. |
| Apple | **Pros:**  - Low signaling latency compared with UP solution (whose termination entity is in LMF or CN or vendor's AI/ML server)  - Native support of multi-vendor interoperation because this solution implies gNB understands and manages the AI/ML model  - The existing RRC signaling solutions can be reused as baseline, at least including delta signaling and segementation.  - SRB transmission is generally more robust than DRB (assuming gNB is not aware of AI/ML model transfer in one DRB as in legacy).  **Cons:**  - The following spec changes may be required:  1) Maximum RRC segementation number may need to increase.   * Whether/how to increase depends on RAN1 input on AI/ML model size per use case, and UE's memory modeling to store AI/ML model(s)   2) Whether to introduce new RRC message or new SRB   * We believe it is stage 3 issue. RAN2 should first focus on stage 2 issue, i.e. discuss and agree requirements and assumptions to use RRC signaling to deliver AI/ML model.   3) Whether to pursue service continuty enhancement (e.g. in handover).   * First, we are not sure whether this is an essential issue, which depends on how frequent the gNB to send new/updated AI/ML to the UE.   + If it is not frequent (e.g. used to deliver offline training outcome), we believe the message/segements lost issue during HO mentioned by some companies is a corner case.   + If it is frequent (e.g. used to deliver inference model in CSI compression), we agree the message/segements lost issue during HO may need to be considered. However, we believe this case may finally have bottleneck in legacy L1 signaling reporting after HO (e.g. after HO, CSI-RS resource needs to reconfigure via RRC, and the UE needs to recalcuate CSI anyway. Then, the new CSI-RS configuration and AI/ML model can be carried in same RRC mesage.). * Secondly, assuming RAN2 agree the enhancement is needed, we think the solution is straght forward with limited spec impacts (i.e. extend loss-less delivery to SRB). More specifically, it can be on top of Rel-17/18 specified QoE continuity for mobility with enhancement for segementation handling (i.e. the UE doesn't need to refresh segements buffer in-between HO).   **Impacts to other WGs:**  - No  **Suitable use case:**  **- T**his model is suitable at least for inference model delivery of CSI compression and beam management whose termination entity is gNB, due to signalig latency consideration.  - Whether it is suitable for training model delivery depends on RAN1 input (e.g. if only offline training is supported in this release, it may not be needed. More discussion is needed if online training is supported).  **Response to some above comments:**  - On model size   * We believe this is RAN1 issue and RAN2 should wait RAN1 input. Actually, RAN1 is discussing to conclude a typical payload size table to send RAN2 in last RAN1 meeting.   - Model duplication may be required at every gNB   * We are not sure what "moldel duplication" means. Different gNB at least need to get timely updated (i.e. different) model parameters for inference purpose of CSI compression and beam management.   - F1 overhead in the split gNB architecture.   * This is RAN3 issue. We are not sure why F1 overhead is essential issue. And even if it is true, we have the same issue in QoE.   - Issues of HO, whether to introduce new RRC message/SRB   * See above comments. For HO, we need more discussion on whether it is essential issue. Whether new RRC message or SRB are stage 3 issue.   - Slow signaling time scale of RRC   * We think LG mentioned issue is LCM issue rather than model transfer. It doesn't prevent to use L1/L2 to activate/deactivate one model if multiple models are configured via RRC. |
| OPPO | Pros   * Less latency compared to Option2-CP soltion.   Cons   * RRC message segmentation number may become extremely large if the AI model size is large enough, but current maximum RRC message segmentation number is 16, more addition, the larger RRC message segmentation number we have the longer time delay we get for model transfer/delivery, so this Option may be applicable to small size AI model transfer/delivery, e.g. several KByte to several MByte; * Extra spec effort, e.g. new SRB, is needed to avoid the AI model transfer/delivery procedure blocking high priority RRC data transmission via SRB1/2/3; * To achieve service continuity for AI model transfer/delivery, HO preparation message may be enhanced to inform the target cell of the AI model transmission status in order to achieve delta AI model transmission at target cell; otherwise, the whole AI model needs to be transmitted again at the target cell; * To achieve service continuity for AI model transfer/delivery in RAN, AI model may be stored at NG-RAN, which costs too much memory at each NG-RAN for the similar AI models compared to Option2-CP solution ; * Model management like model update and model sharing procedure is challenging when AI models are stored at each NG-RAN.   Extra comments:   * AI model transmission authorization/registration procedure may be needed before model transfer/delivery, this may involve SA2 work and common for all 3GPP-based solutions; |
| Dell Technologies | We mainly agree with QC on the potential issues of the CP AI/ML model delivery/transfer  **Pros**:   * Clearly less latency compared to CP Option 2 and UP options * Lesser spec impact, and is limited to RRC specs * AI/Ml can be treated as part of UE context, depending on UE capability.   **Cons**:   * Only support small sizedAI/ML models; however, the number of RRC message segementations is significantly expaneded. * May potentially block other non-AI/ML mode RRC information if the a larger AI/ML model is exchanged as part of existing RRCreconfigurations, specially in HO and poor RF conditions scenarios. * Likely need a new RRC message definition to void the above mentioned RRC related issues, and which can be eaisely tuned and adaptively configured to AI/ML model sizes and nature, more than existing RRC signaling. |
| Huawei, HiSilicon | Firstly, for all possible solutions, we think it is required to have common evaluation metrics. Here are our views:   * Model characteristics, such as model size, model transmission/update frequency, transmission latency * Overhead * Impacts due to handover * Impacts to RAN2 and perhaps other WGs * Inter-operability impacts   Secondly, for the above metrics for all possible solutions, there may be some categories for them, e.g. for model size, there may be large model and small model; for transmission latency, there may be low/high latency requirements. In our opinion, it will be helpful for RAN2 to check the metrics/combinations and then better compare the solutions.  Here are our views on Option 1 – CP solution:  **Pros:**  - It can be flexible. For model characteristics, if model size is large without strict transmission requirements, we can think about SRB4 like solution (a low priority SRB); if model size is small with strict transmission requirements, existing SRB can work  - Existing RRC signalling can be considered as baseline, and enhancements can be discussed  **Cons:**  - Overhead:   * Between this CP solution and other UP solutions, the main difference on overhead is at RRC layer, and the overhead below RRC (PDCP/RLC/MAC/PHY) should be similar. At RRC layer, if there are some RRC segments for the model transfer/delivery RRC message, it may introduce some overhead   **Other comments:**  - Impacts due to handover:   * As some companies mentioned above, an incomplete control plane model transfer has to be restarted upon mobility, so it may lead to some problems. Our view is that firstly we may need to check whether it is required to continue model transfer/delivery during HO, and secondly, it depends on model characteristics, e.g. if mode size is small, the issue may not happen frequently   - For impacts to RAN2, there may be the following changes:   * Maximum number of RRC segmentation number may need to be enhanced * Re-use existing RRC message or introduce new RRC message * Re-use existing SRB or introduce new SRB |
| NEC | The benefit of RRC based solution should be that gNB can take the control of the AIML model transfer itself, which can not be achieved by traditional UP based solution.    We think the main issue of RRC based solution is that it may need to extend the existing RRC segmentation support, since the current RRC segments for RRC message is limited. However this does not present a block issue from the perspective of message transmission.  Meanwhile, we see that in legacy system design there is well known handling for data transmission interruption but not for signlling transmission interruption, due to e.g. radio link failure. It would be appreciated to discuss how to support the RRC message based AIML model transfer in the case of signlling transmission interruption. We do not think this is a block issue for RRC message based AIML model transfer. But RAN2 should take such discussion. |
| Lenovo | Pros:   * Existing RRC procedures/principles can be used with enhancement, e.g., SRB concept, delta configuration. * Higher priority and less latency compared to AI model transfer/delivery via UP/DRB   Cons:   * Maximum number of RRC segmentations needs to be extended. But still the maximum AI model size is upbounded by a certain value. * In case of any radio link problem, e.g., RLF, whether/how to resume the AI model transfer/delivery is unclear.   Some aspects mentioned by above companies are dependent on if the AI model is trained at gNB or trained by OAM/CN and transferred/delivered to gNB. For example, if RAN2 needs to specify AI models (at least some parameters) in RRC spec, and if model duplication is needed between gNBs (e.g., during handover source gNB needs to transfer/deliver the AI model in use to the target gNB).   * + If the AI model is trained by each gNB, then it is true that RAN2 needs to specify AI models in RRC spec (which could be tricky), and source gNB probably needs to transfer/deliver the AI models in use to the target gNB during handover (which could be signaling heavy).   + In some other possible deployments, AI model duplication between gNBs is not needed and RAN2 does not need to specify AI models. In one example, AI models could be trained by OAM and delivered to all gNBs in the same network. In that case, when gNB transfers/delivers the AI model to UE, a AI model could be conveyed in a RRC container. Also, during handover, source gNB could indicate target gNB about the AI model in use using some indicator instead of transferring the complete AI model, since they have obtained the same AI model from OAM. |
| Spreadtrum | Pros:   * Low latency compared with CN-CP based solution especially when model size is small; * Beneficial for the support of RAN AIML function, i.e., RAN AIML model (e.g., CSI compression and prediction, Beam management) can be stored and managed by RAN itself;   Cons：   * RRC segmentation number may large considering huge AIML model size, which may cause extra delay; * It may be complex during HO, as mentioned by companies, the whole AIML model needs to be retransmitted in target gNB if model transfer/delivery is not completed before handover; * Using RRC signalling (e.g., RRCReconfiguration message) to carry AIML model may block other control plane info. Thus, using existing RRC signalling or define a new RRC signalling should be discussed. |
| CMCC | Pros:  - No multi-vendor interoperability issue  - Spec impacts are limited in RRC without any impacts on other WGs  - Less signaling latency compared to other CP options and UP solutions  Cons:  - Difficult to support a large size model by RRC signaling due to limited RRC size, and RRC segmentation number may need to be extended which depends on RAN1 input on model size per use case  - The signaling overhead may be heavy  - RAN2 may need to specify AI/ML model (e.g. model structure, model parameters) in RRC which have large spec impacts |
| Nokia, Nokia Shanghai Bell | Pros:  - Limited specification impact for supporting transfer/delivery of a model with a few KB in size.  - Additional security and verification may not be necessary as the UE already established security before the transfer is initiated.  - Attached metadata to the transfer/delivery process is synchronized with the transfer/delivery process.  - Less latency compared to other solutions (CP-Option2 and UP).  Cons:  - Cannot support large models (>45KB) without increasing the maximum allowed RRC segment limit past 16.  - Model transfer/delivery failure and recovery is not clear (e.g., during handover).  - RRC message segmentation could cause handover failures and, if the model transfer/delivery is not aborted, could cause connection reestablishment failure. |
| ZTE | Pros:   * Limited specification impact, the exsited tunnel for transferring the RRC control signaling can be reused for AI model transfer. * No security issue.   Cons:   * HO will casue the model transfer data loss. * If the SRB for transferring the AI model is shared with other RRC procedures, may cause the RRC procedure delay, the request of a new SRB dedicatedly used for AI model tranfer. * For UL, the PDCP duplication for SRB cannot be dynamically controlled. |
| Ericsson | While working on an RRC-based solutions seems to be the “straightforward” one from a RAN2 standpoint (especially given the pros highlighted). We are mostly worried about the cons that have been stressed by other companies above. We are of course open to discuss methods to address those. However, it could simply become unfeasible to carry the (arguably) large models with RRC. Our proposal would then be to wait for RAN1 input which could simplify/clarify the way forward. |
| Fujitsu | Pros:   * Seamlessly support light model updating without big spec impact, can be deployed in the near future for some applications for commercialization. * Have quick response and low latency, which is suitable for the model with bad generalization performance and need quick updating frequently. * Friendly to flexible model transfer due to the transferred contents can be seen and managed in RRC messages, so customized partial transmission is available.   Cons:   * When the model size is large and the number of requested model transfer is large, the RRC resources may be occupied by AI model transfer only. * HO issues. |
| CATT | Pros:   * To transfer/deliver the AI/ML model via RRC signalling means that the model content is carried on SRB, which we think is an appropriate method used between UE and gNB; * It is possible that the model is visible to the RRC layer, thus it is convenient to control the model update, e.g. to have delta-configuration to reduce the signalling size;   Cons:   * When the model size is big, segmentation is necessary, but current RRC spec only allows 5 segments for DL and 16 segments for UL which can not fulfil the AI/ML model transfer requirement. More segments may be needed for AI/ML model transfer/delivery; * New SRB type may be defined which may introduce more complexity, and the priority of the new SRB should be considered. |
| Futurewei | Pros:   * Shorter model transfer/delivery latency; * Lower signaling overhead (fewer inter-layer signalling comparing to Option 2); * Fewer multi-vendor issues (this option assumes gNB manages the models); * gNB has the control and flexibility of choosing the right model and put it to work at the right moment; * Model transfer/delivery will have higher priority and more robust.   Cons:   * Difficult to transfer/deliver large-size models due to RRC segmentation limit; * The RRC segmentation limit could in turn cause the reflush of the security key, causing extra delay; * (Heavy) model transfer/delivery traffic adds burden to the CP and could block other high-priority control messages; * Although it may not happen frequently, model transfer/delivery during mobility or RLF may need to be retransmitted (the retransmission mechanism is not clear); * This approach implies that gNB needs to host the functions of model management (including model transfer and delivery); although doable, it adds burden to gNB; * May need new/enhanced RRC signaling (e.g., new SRB category). |
| China Unicom | Pros:   * Low latency, which can serve the real-time service for optimization; * High reliability, which are beneficial for the vertical industries. * Good network controllability, e.g. Modely updated easily.   Cons:   * For real-time service, it seems that the model size is limited even with RRC segmentation mechanism; * It will consume lots of signalling overhead if the model transfer/delivery frequently;   Have big impacts on the specs, e.g. support the model transfer/delivery in mobility scenario. |
| Mediatek | Agree with HW that we should list and discuss the common evaluation matrics first. From our side, following aspects can be considered:   * Capability to transfer model of large size * Latency of model transfer * Support of model/model parameter update during mobility * Signaling overhead * Specification impact * Applicability   Pros:   * Lower latency of model transfer compared with option 2, 3, 4. * Dela configuration is possible, when only model parameter update is required. * Compatible with current mobility procedure and provide the possibility to update model during mobility * Limited specification impact.   Cons   * gNB coordination is required during UE mobility for model/parameter update * Hard to support large model size * More issues need to be addressed if CP solution is used to transfer models of large size, e.g. latency, overhead, increased number of segments.   Option 1-CP solution is favor for model transfer with small model size.  Extra comment:  Considering the generalization performance of the AI/ML model, the AI/ML model may be applicable to a cell or a site. Model may or may not change as UE moves to different cells/sits. The coordination between source gNB and target gNB is required to sync the model information. If model doesn’t change, the source gNB may need to forward the model information to the target gNB. If model changes, model/model parameter update during mobility needs to be considered. But if model transfer is hard to be supported during HO, the network can disable AI/ML operation before HO and start model transfer after HO. The drawback is the long latency of model availability in the target cell.  For interoperability, I am wondering whether it’s more impacted by the model format instead of model transfer method? If it is proprietary formation, the interoperability may be concerned. |
| Samsung | **Pros:**   * Lower latency and/or specification impact compared with other options.   **Cons:**   * Transfer/Delivery of large size models may impact RRC signalling. However, one potential solution could be to define a new SRB type to deliver AI/ML model.   We also agree with OPPO’s view/extra comment, that a model registration (and/or authorisation) procedure, with the network, may be needed before the transfer/delivery of the model. |

#### 2.2.2.2 Option 2 – CP solution

For this CP solution, the principle is that CN (except LMF) can transfer/deliver AI/ML model(s) to UE via NAS signalling. As discussed in phase 1, the applicable use cases are CSI feedback enhancement and Beam management, and there are no specific considerations for Positioning accuracy enhancement.

The basic flow for this CP solution is shown in figure 2 below.



Figure 2: Basic flow for Option 2 – CP solution

**Q7: Regarding Option 2 – CP solution, do companies agree with the principle and the basic flow described above?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
| vivo | Yes | Figure 2 can be a baseline, and whether more procedures (e.g., response) are needed can be decided by SA2 if this option is concluded to be supported. |
| Xiaomi | Yes |  |
| Qualcomm | May need additional procedures/Signaling | Currently, the NAS signaling is between UE and AMF. However, AMF may not aware of the ML model, it requires more procedures to involve other 5GC entities. The proposed procedure may imply a misunderstanding that the CP solution only impacts the NAS signaling between UE and AMF. It is proposed to add a note that the procedures between different 5GC entities may be required and need SA2’s study. |
| LGE | Yes | The Figure 2 is considered incomplete given that there may exist ML-dedicated Network Function that can interface with CN. |
| Apple | Yes with comments | We agree that Figure 2 is baseline, but we are wondering whether below two cases can be included in Option 2:  1) Termination entity is OAM which is captured in RAN3 TR.  2) Termination entity is OTT server, i.e. QoE like solution (AI/ML model is from application layer but delivered to UE via SRB) |
| OPPO | Yes with comments | Similar comments as Q5, AI model transmission authorization/registration procedure may be needed before model transfer/delivery and both network triggered and UE triggered model transfer/delivery should be considered, only network triggered model transfer/delivery is reflected into the Figure 2, better to clarify this also. |
| Dell Technologies | Yes with comments | Require new NAS signaling/SRBs since AMF is not necessary te entity handling the AI/ML model delivery/transfer. However, we are OK with Fig 2 as a baseline. |
| Huawei, HiSilicon | Yes |  |
| NEC | See the comments | We share the same understanding with Qualcomm on that the AMF itself may not hold the AIML model. We should just assume that there is a 5GC entity holding the model, which would present the need to have 5GC internal signalling/data transfer to enable option 2. |
| Lenovo | Yes with comment | Yes in general. Agree with QC’s comment that currently NAS signaling is between UE and AMF. Which CN nodes are involved may need further discussion. |
| Spreadtrum | Yes with comments | Figure 2 can be the baseline for CN-CP solution.  Other signalling enhancement like pre-signalling of authorization, capability indication etc, and subsequent response signalling can be further discussed.  And CN is a wide concept which may include several entities like AMF, PCF, NWDAF etc. The potential internal interaction among CN entities may exist and should be considered by SA2. |
| CMCC | Yes | We are fine with figure 2 as baseline. Which entity performs model transfer or procedures between different 5GC entities needs SA2 study. |
| Nokia, Nokia Shanghai Bell | Yes | The basic flow is fine. Please see Q8 for further details. |
| ZTE | Yes, but see comments | The basic flow is also okay to us.  However, even the tunnel (i.e. SRB+NG-C) for AI model transfer have been there, whether there is impact on 5GC is still unclear since the current AMF function does not include the AI model handling, SA2 shall be involved for this option to discuss whether to extend the current AMF function or just add an OAM associated with the AMF.  By the way, we are wondering what use cases for which the correponding model transfer need to be terminated between UE and AMF….? |
| Ericsson | No | Check our input to Q8. |
| CATT | Yes with comments | In this option, the AI model is transferred from CN node to the UE. It can be considered as one of the model transfer option between UE and network node. One more comment is that at least in current RAN1 discussion and assumption, two use cases of CSI feedback enhancement and Beam management only involve UE and gNB. The previous comments seem to assume the model is already avaible in CN. However, the model transfer from gNB to the CN node is necessary to make the whole procedure compelte. For Option 1 there is no such issue as RRC singaling terminates at gNB. |
| Futurewei | Yes | Figure 2 can be the baseline. |
| China Unicom | Yes | We agree with the basic flow, but how to invole SA2 need to be discussed. |
| Medaitek | Yes | It’s ok to have Figure 2 as baseline. The only reference point between UE and CN is N1(NAS) interface. RAN2 doesn’t need to figure out the interactions of NFs in 5GC, which can be left to SA2. |
| Samsung | Yes with comments | Figure 2 is generally ok, as a baseline for Option 2.  RAN2 may need input from SA2 on potential impact(s) of Option 2 on other network entities. |

For this CP solution, it is suggested to collect pros/cons and other comments.

**Q8: Please provide your comments on Option 2 – CP solution in the table below, such as pros/cons, impacts due to model size/latency, use case specific analysis/comments.**

|  |  |
| --- | --- |
| **Company** | **Comments** |
| vivo | **Pros:**   * No multi-vendor interoperability issue if the model format/description is clarified during model transfer. * The NAS layer can be responsible for segmentation. Thus no impact on the RRC layer. * If the model is visible to the NAS layer, delta configuration can be used to reduce the signaling overhead. * If the CN entity can subscribe to the handover event from AMF, suitable models can be updated/transferred to UE with limited latency.   **Cons:**   * As the NAS signaling is transparent to gNB, SRB2 will be reused for model transfer, which may introduce a significant impact on the existing signaling transmission (e.g, Registration procedures) if the model size is huge. * Significant SA2 impact, e.g., they need to study which CN functionality is responsible for the model transfer. |
| Xiaomi | Pros:   * May be able to deliver large size model, up to SA2 evaluation. * May require less signalling exchange during handover, if AI model is applicable in multiple cells   Cons:   * Large latency compared with gNB solution. * Require more inter-layer signalling compared with gNB solution, since the use case is mainly located in physical. * Not clear whether CN node is able to determine the applicable AI for physical use case   However, it’s better leave to SA2 to evaluate the feasibility and Pros/Cons. We suggest RAN2 to focus on other solution. |
| Qualcomm | Cons:   * it requires SA2 support, it is not clear which 5GC entity can provide the RAN level ML model, it requires 5GC to understand the RAN level ML model. * It is not clear how to standardize the ML model via NAS signaling. * It will increase the huge signaling burden since the ML model size is always very high. |
| LGE | Pros:   * Model transfer/delivery is almost transparent to RAN2 spec.   Cons:   * Since RRC needs to carry the NAS message including ML model, it also suffers from the same issue of a big message size as RRC-based CP delivery.   The feasibility of this solution should be evaluated by SA2, rather than by RAN2. |
| Apple | **Pros:**  - Native support of multi-vendor interoperation because this solution implies gNB understands and manages the AI/ML model  - Limited or minor RAN2 spec impact  - SRB transmission is generally more robust than DRB (assuming gNB is not aware of AI/ML model transfer in one DRB as in legacy).  **Cons:**  - The signaling latency is higher than option 1 (i.e. gNB solution)  **Impacts to other WGs:**  - SA2 and CT1, including:   * 5GC entity to manage AI/ML model * NAS signaling change to include AI/ML model.   **Suitable use case:**  - Transfer of offline training AI/ML model  - It doesn't make sense to let CN aware of inference model for CSI compression and beam management. |
| OPPO | Pros:   * Model management like model update and model sharing procedure is easy compared to Option1-CP solution; * Service continuity on model transfer/delivery is easy to achieve compared to Option1-CP solution; * NAS spec will handle the model data segmentation issue, which means less impact on AS spec.   Cons:   * Option2-CP solution has larger time delay than Option1-CP solution; * Extra spec effort, e.g. new SRB, is needed to avoid the AI model transfer/delivery procedure blocking high priority NAS data transmission via SRB2;   Extra comments:   * AI model transmission authorization/registration procedure may be needed before model transfer/delivery, this may involve SA2 work and common for all 3GPP-based solutions; |
| Dell Technologies | **Pros:**   * No or little RAN2 spec impact. * Similar pros to CP option 1   **Cons:**   * Higher latency than CP option 1 * New SRB and NAS signaling specifications may be needed since (1) AMF is not (always) handling AI/ML model delivery/transfer over NAS, and (2) not block other non-AI NAS critical information. |
| Huawei, HiSilicon | In Q6, we comment that there should be common evaluation metrics for all possible solutions. Here are our views on Option 2 – CP solution:  **Pros:**  - Similar to Option 1 – CP solution, it can be flexible. For model characteristics, if model size is large without strict transmission requirements, we can think about SRB4 like solution; if model size is small with strict transmission requirements, existing SRB can work  - Existing NAS signalling can be considered as baseline, and enhancements can be discussed  **Cons:**  - Overhead:   * Between this CP solution and Option 1 – CP solution, one question is whether NAS signalling should fit the RRC message size requirement or not. If yes, we think NAS needs to do the segmentation, and then it may introduce some overhead   **Other comments:**  - Impacts due to handover: same as our response to Q6  - For impacts to RAN2, we think that SA2/CT1 can check details and then RAN2 can check possible RAN2 impacts |
| NEC | We do not think NAS based solution is a practical solution, even though for completeness it is ok to be listed by rapp here, since at least it highly depends on the fact that the 5GC should hold the AIML model. However if 5GC holds the model, why UP based transmission is not used, which is the traditional way to transmit the data between UE and 5GC.  This key disadvantage of this option is the need to take segmentation for NAS message, which further need to have underlined support from SRB. As a consequence, all of the disadvantage of option 1 may hold for this option also. In addition, this option may require lots of SA2 work, which should be discussed by SA2. |
| Lenovo | Pros:   * Less RRC impacts * Compared to UP/DRB solution, AI model transfer/delivery via SRB has higher priority and less latency   Cons:   * SA2 impacts are expected * This option implies the AI model could be trained by CN node. In that case, how does a CN node obtain all necessary training data (e.g., L1/L3 RAN measurements) is tricky. In legacy, the exposure of RAN measurements to CN is quite limited. Besides, CN is not a good option for later on model monitoring/activation/deactivation/fallback/update that requires less latency * The model transfer/delivery is transparent to gNB, it could be tricky to get gNB involved in the AI model LCM. It could be problematic when the network needs to be in control of what happening at the UE side and especially in two-sided models where one side of the model is intended to be located at the network side.   In case of any radio link problem, e.g., RLF, whether/how to resume the AI model transfer/delivery is unclear. |
| Spreadtrum | Pros:   * Little RAN2 impact; * Benificial for contiunity of model transfer/delivery in mobility scenario. Target gNB and source gNB may connect to the same CN entities which can provide the AIML model.   Cons：   * It is unclear that whether CN entities can manage RAN AIML model, i.e., CSI compression and prediction, beam management. * Larger latency than option1-CP based solution. * NAS signalling is carried in RRC signalling. Thus it will have the similar cons with option1- CP based solution, e.g., segmentation, new SRB consideration. |
| CMCC | Pros:  - No multi-vendor interoperability issue  - No/less spec impacts on RRC which is transparent for RAN2  Cons:  - Significant impacts on SA2/CT1, however, we are not sure that whether it is suitable for CN to determine AI/ML model for physical use case, it requires more SA2 evaluation  - Higher latency compared to option 1  - The signaling overhead may be heavy |
| Nokia, Nokia Shanghai Bell | Pros:   * Model transfer/delivery is transparent to gNB.   Cons:  - Support for large model is limited since NAS signalling will be encapsulated in SRB2.  Therefore, similar cons of RRC signalling apply here.   * Additional signalling could be required for interoperability between the UE and gNB.   Additional comment: SA2 evaluation may be needed for further analysis. |
| ZTE | PROs:   * Limited RAN spec impact   CONs:   * Since the SRB is also used for this option, so have the similar CONs in option 1. * SA2 shall be involved, but there is no TU allocated to SA2 for disucssion. * No obvious use case which can apply the model transfer terminated between UE and AMF. |
| Ericsson | On the one hand, and as stressed by others above, the situation for this case does not seem to be considerably different to Option 1, as perhaps large NAS messages would need to be carried over RRC.  However, what worries us the most is that Option 2 involves other WGs without TUs allocated to this SI. For that reason, our proposal would be to consider this option as out of RAN2 scope. |
| Fujitsu | Pros:   * As NAS signaling cannot be touched by RRC container and be transparent to RAN2, so there will be few specification impacts. * Latency may be acceptable for some applications compared to UP solutions.   Cons:   * The motivation for using the non-LMF blocks in CN (e.g., AMF) to transfer model to UE is unclear, currently, there is no need for CN to deploy or re-train AI/ML model for the agreed 3 use cases. * This solution will be tricky for the application of two-sided models. * May introduce extra workload with other WGs. |
| CATT | Pros:   * Less RRC impact, the segmentation can be performed in NAS spec and decided by SA2.   Cons:   * The latency may be increased compared with CP based solution option 1. |
| Futurewei | Pros:   * Able to handle larger-size model without the RRC segmentation limit issue * Easier for centralized model management; * Less burden to gNB; * Like Option 1, multi-vendor interoperability is not/less an issue; * Less impact to RAN2;   Cons:   * Even though RRC segmentation is no longer an issue, since NAS is carried by RRC, large-size model transfer/delivery could still cause CP-related issues, such as CP burden and blockage of other high-priority control messages; * Longer latency comparing to Option 1, but may not be a big issue as model transfer/deliver may not be delay sensitive; * Standard impacts to SA2 (requires SA2 support); * More inter-layer signaling comparing to Option 1; * gNB has no control to the LCM of the model, which could be problematic.   Additional functions needed at CN/AMF   * The capability to host and manage the models at 5GC (current NAS signaling is between UE and AMF);   Other comment:   * This solution assumes the CN manages the models. In this case, the UP-based solution is the more natural solution. |
| China Unicom | Pros:   * Less RAN2 inpact.   Cons:  Not feasible for the real-time AI service. |
| Mediatek | **Pros:**   * Dela configuration may be possible, when only model parameter update is required. * Easier to support model/model parameter update during UE mobility, since the gNB can require the model information from CN * Less RAN2 specification impact   **Cons**   * Longer latency of model transfer compared with option 1. * Hard to support large model size * Requires SA2 involvement and evaluation   Extra comment:  We tend to agree that model transfer from CN may not be proper for the use cases of AI/ML operation purely over air interface, e.g. for CSI and BM, requiring RAN to be responsible for the life cycle management. How to make RAN node be aware of AI/ML model needs to be considered further. |
| Samsung | **Pros:**   * Possibly no or minor impact to RAN2 spec.   **Cons:**   * Latency could be higher that Option 1 - CP. * Transfer/Delivery of large size models may still have impact on RRC signalling (i.e. carrying NAS message that contain AI/ML model). * Impact to other WGs, for example, SA2, as it is not clear at this stage which 5GC entity may be impacted by this Option. So input from SA2 would be needed if RAN2 agree to adopt Option 2 - CP. * Impact to NAS signalling. So input from CT1 may be needed on this point.   We also agree with OPPO’s view/extra comment, that a model registration (and/or authorisation) procedure, with the network, may be needed before the transfer/delivery of the model. |

#### 2.2.2.3 Option 3 – CP solution

For this CP solution, the principle is that LMF can transfer/deliver AI/ML model(s) to UE via LPP signalling. As discussed in phase 1, the applicable use case is Positioning accuracy enhancement.

The basic flow for this CP solution is shown in figure 3 below.



Figure 3: Basic flow for Option 1 – CP solution

**Q9: Regarding Option 3 – CP solution, do companies agree with the principle and the basic flow described above?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
| vivo | Yes with comments | Figure 1 can be a baseline and the following procedures can be further discussed:   * The model transfer/delivery can be a procedure with a response. The response can be used for the UE to indicate the model is ready for inference/activation when UE has re-segmented/compiled the model. * The UE may request model transfer/delivery beforehand based on its capability and preference. Taking PRS configuration as a reference, the UE may request the model before the LMF delivers the model to the UE. |
| Xiaomi | Yes | The figure 3 is the principal and detailed signalling procedures can be further discussed. |
| Qualcomm | No | We believe this should be the sub-use case of option 2. Further details should be left to SA2 on whether a unified model delivery/transfer method is required, or a separate model delivery/transfer method is desired for positioning. |
| Apple | Yes | We think that whether LMF (i.e. option 3) or a new 5GC entity (i.e. option 2) to manage AI/ML model should be left to SA2. |
| OPPO | Yes with comments | Similar comments as Q5, AI model transmission authorization/registration procedure may be needed before model transfer/delivery and both network triggered and UE triggered model transfer/delivery should be considered, only network triggered model transfer/delivery is reflected into the Figure 3, better to clarify this also. |
| Dell Technologies | No | Agree with QC, we think this option can be part of CP option 2, and leave it to SA2 to determine the 5GC entity handling AI/ML model delivery/transfer. |
| Huawei, HiSilicon | Yes |  |
| NEC | Yes | The figure may serve as stage 2 concept |
| Lenovo | Yes | It seems straight forward to us that LMF is responsible of AI model provision to UE. We can take it as the start point unless SA2 states otherwise. |
| Spreadtrum | Yes | The figure 3 can be the baseline for positioning use case.  Whether it is LMF who provides AIML model directly or LMF gets AIML model from other entities in 5GC (e.g., NWDAF) can leave to SA2 discussion. |
| CMCC | Yes |  |
| Nokia, Nokia Shanghai Bell | Yes | The basic flow is Ok. The detailed signalling can be discussed further. |
| ZTE | Yes |  |
| Ericsson | No | Check our input to Q10. |
| Fujitsu | Yes | Similar reason as Q5, and this option is more natural than option 2, because the LMF has the ability to collect data from multiple gNBs and UEs for the model training and updating. |
| CATT | Yes |  |
| Futurewei | Yes | This can be the baseline for Option3. |
| China Unicom | Yes | We are fine with the basic flow. |
| Mediatek |  | Agree with Qualcomm that option 3 is a sub use case of option2. The pros and cons analysis share the same logic and methodology for both option 2 and option3. |
| Samsung | Yes |  |

For this CP solution, it is suggested to collect pros/cons and other comments.

**Q10: Please provide your comments on Option 3 – CP solution in the table below, such as pros/cons, impacts due to model size/latency, use case specific analysis/comments.**

|  |  |
| --- | --- |
| **Company** | **Comments** |
| vivo | **Pros:**   * No multi-vendor interoperability issue if the model format/description is clarified during model transfer. * The LPP layer can be responsible for segmentation. That is, no impact on the RRC layer. * If the model is visible to the LPP layer, delta configuration can be used to reduce the signaling overhead. * As LMF can subscribe to the handover event from AMF, suitable models can be updated/transferred to UE with limited latency. * Compared to Option 2 (i.e., CP solution between other CN entities and UE), less SA2 impact is foreseen as option 3 reuses LMF and LPP to handle the model transfer.   **Cons:**   * As the LPP signaling is transparent to gNB, SRB2 will be reused for model transfer, which may introduce a significant impact on the existing signaling transmission (e.g, Registration procedures) if the model size is huge. |
| * Xiaomi | Pros:   * LMF has a large amount of UE location data and is feasible for model training and delivery. * The LPP supports the segmentation and the larger size model can be delivered; * The LPP can be delivered by the user plane and thus means the UP –based solutions also can be supported by the LPP enhancement * Limited AS layer impact   Cons:   * May not support the model transfer/delivery in uplink |
| Qualcomm | Same as answers to Q7, Q8, and Q9. |
| Apple | **Pros:**  - Native support of multi-vendor interoperation because this solution implies gNB understands and manages the AI/ML model  - Limited or minor RAN2 spec impact  - Native / better support of AI/ML based positioning (entity for AI/ML model management and positioning management is co-located and LPP signaling can be reused).  - SRB transmission is generally more robust than DRB (assuming gNB is not aware of AI/ML model transfer in one DRB as in legacy).  **Cons:**  - The signaling latency is higher than option 1 (i.e. gNB solution)  **Impacts to other WGs:**  - SA2 and CT1, including LPP signaling change to include AI/ML model.  **Suitable use case:**  - AI/ML based Positioning if SA2 agree LMF to manage AI/ML model for positioning.  - It doesn't make sense to let CN aware of inference model for CSI compression and beam management. |
| OPPO | Pros:   * Model management like model update and model sharing procedure is easy compared to Option1-CP solution;   Cons:   * Option3-CP solution has larger time delay than Option1-CP solution; * Extra spec effort, e.g. new SRB, is needed to avoid the AI model transfer/delivery procedure blocking high priority LPP data transmission via SRB2; * This option may only be applicable to positioning use case.   Extra comments:   * AI model transmission authorization/registration procedure may be needed before model transfer/delivery, this may involve SA2 work and common for all 3GPP-based solutions; |
| Dell Technologies | Same as Q8 |
| Huawei, HiSilicon | In Q6, we comment that there should be common evaluation metrics for all possible solutions. Here are our views on Option 3 – CP solution:  **Pros:**  - Similar to Option 1 – CP solution, it can be flexible. For model characteristics, if model size is large without strict transmission requirements, we can think about SRB4 like solution; if model size is small with strict transmission requirements, existing SRB can work  - Existing LPP signalling can be considered as baseline, and enhancements can be discussed  **Cons:**  - Overhead:   * Between this CP solution and Option 1 – CP solution, one question is whether LPP signalling should fit the RRC message size requirement or not. If yes, we think LPP need to do the segmentation, and then it may introduce some overhead   **Other comments:**  - Impacts due to handover: same as our response to Q6  - For impacts to RAN2, there may be the following changes:   * May need to support the segmentation of LPP message * Re-use existing LPP message or introduce new LPP message |
| NEC | This mechanism requires SA2 discussion, since basically the positioning server is hosted in the 5GC domain.  This advantage of this option is that the positioning server is already available, so then AIML model can be seen as additional (special assitant) data for positioning from data transmission perspective. |
| Lenovo | Pros:   * For direct AI/ML positioning, legacy LPP procedures/principles can be used with possible enhancements for LMF to collect training data and train the AI model properly. * Segmentation at LPP layer is supported * Less RRC impact * Compared to UP/DRB solution, AI model transfer/delivery via SRB has higher priority and less latency   Cons:   * Whether/how to specify the AI models (at least some parameters) in LPP spec could be tricky. |
| Spreadtrum | Pros:   * Little RAN2 impact; * It is suitable for LMF to be responsible for positioning AIML model transfer/deliver;   Cons:   * Larger latency than option1-CP based solution. * LPP signalling is carried in RRC signalling. Thus it will have the similar cons with option1- CP based solution, e.g., segmentation, new SRB consideration. |
| CMCC | Pros:  - No multi-vendor interoperability issue  - Limited impact on RRC  - It’s specific and applicable for AI/ML based positioning use case, LPP signaling can be reused  Cons:  - Higher latency compared to option 1  - The signaling overhead may be heavy  - Need more work in SA2/CT1 |
| Nokia, Nokia Shanghai Bell | Pros:  - The existing LPP protocol can be extended to support model transfer/delivery as long as the model size fits within the limitations of the LPP message size and number of segments possible.  - The CP method would ensure a synchronization between the LMF and UE state with regard to available models.  Cons:   * Limitation on model size. Until it is known what to expect for model size, it will be difficult to determine whether CP LPP can support model transfer/delivery. |
| ZTE | Cons/Pros are similar as Q6, Q8 since the SRB is still used for model transfer. |
| Ericsson | As commented in Option 2, RAN2 work will depend on other WGs to progress, for which we prefer to leave this option of scope of this SI. |
| Fujitsu | Pros:  Similar to Q8  The existing LPP specification has complete framework to support both model transfer/delivery and capability report/model registration for AI/ML positioning models.  Use LPP is better than RRC by reusing other positioning messages carried by LPP signals.  Cons:  The latency may cause problems, for positioning applications, the latency will be a big problem because the model is more sensitive especially when UE is moving. |
| CATT | Pros:   * The LPP spec is easy to extend to support AI model transfer between UE and LMF; * Almost no RRC spec impact. |
| Futurewei | Pros:   * LMF is a better place than gNB for positioning use case as the positioning service function and data may already be in the CN. * Other advantages are similar to Option 2.   Cons:   * Disadvantages are similar to Option 2.   Other comments:   * Like Option 2, this case is better handled in the UP. |
| China Unicom | Pros:   * Less RAN2 inpact. * LPP signalling can be resued.   Cons:   * Limited model size.   Larger latency than option1-CP solution. |
| Mediatek | Same as Q8  It’s specific and applicable for positioning. |
| Samsung | See answer to Q8 |

### 2.2.3 UP-based solutions

Based on the outcome of phase 1 discussion, this part is to collect companies’ views on principles, basic flows, pros/cons, and others for each possible UP-based solution. RAN2 impacts and cross-WG impacts can also be discussed. The following solutions have been identified for the RAN2 study due to phase 1 summary:

* Option 1 – UP solution that gNB can transfer/deliver AI/ML model(s) to UE via UP data.
* Option 2 – UP solution that CN can transfer/deliver AI/ML model(s) to UE via UP data.
* Option 3 – UP solution that LMF can transfer/deliver AI/ML model(s) to UE via UP data.

#### 2.2.3.1 Option 1 – UP solution

As discussed in phase 1, for this UP solution, the applicable use cases are CSI feedback enhancement and Beam management, and there are no specific considerations for Positioning accuracy enhancement.

For this UP solution, how it works is not clear to some companies in phase 1 discussion. So it is suggested to collect companies’ views on the principle and the basic flow.

**Q11: Regarding Option 1 – UP solution, please companies provide your views on the principle and the basic flow in the table below.**

|  |  |
| --- | --- |
| **Company** | **Comments** |
| vivo | The User Plane Protocol Stacks specified in 23.501 is copied as follows:    Figure 8.3.1-1: User Plane Protocol Stack  If a new User Plane date terminated at gNB is introduced, it seems to break away from the current UP protocol stack as the legacy UP data is not terminated at gNB and will be further delivered to UPF.  Besides, the proponents are suggested to clarify the difference between this option and introducing a new SRB type. |
| Xiaomi | Existing user plane is not applicable, since it terminates at UE and UPF.  A new AI layer may be introduced to handle the AI model transfer functionality. Both uplink and downlink signalling can be introduced. The protocol could be as following, |
| Qualcomm | In our understanding, the application function (AF) hosting the AI/Mldels and UPF can be collocated with the RAN. We believe that the protocol stack can be implemented at any network entity. The same design principle can be used for any User Plane based model delivery/transfer solution (including option 4). For example, to reduce the delay in the model delivery/transfer I-UPF and AF can be implemented at gNB or can be placed closed to the gNB. |
| LGE | We have a similar view with QC. From architectural point of view, the suggested UP options are not mutually exclusive ones but just implementation variants of a common UP-based architecture that employes a network entity such as Application Function in charge of ML model provisioning. If network deploy the network entity to be collocated with Gnb (as similar to MEC), it is O1, and if not, it is O2.  SA2 should be involved to make any progress with this option. |
| Apple | We agree with vivo and Xiaomi's comments that existing user plane solution doesn't support Option 1. Whether enhancement is required to make option 1 work should be left to SA2. And because it may change very basic protocol stack, we believe RAN2 should not make any conclusion / suggestion / assumption on this solution. |
| OPPO | For this Option, the signalling flow can be illustrated like the following:    Figure x Basic flow for Option 1 – UP solution  For this Option, we think AI models are stored at gNB, and gNB will transfer/deliver AI model to UE directly. |
| Dell Technologies | Similar view to QC and LGE. RAN2 should not conclude on this, and is left to SA2. In our understanding, presented UP architectures are not execlusive. The gNB can host or be collocated with UPF and/or AF depending on the deployment variant. Therefore, there is no need for UP, specially terminated at the gNB. |
| Huawei, HiSilicon | We agree with vivo and Xiaomi's comments that existing user plane solution doesn't support Option 1, and then how it works needs to be discussed first. |
| NEC | We would suggest to reword option 1 as “gNB can transfer/deliver AI/ML model(s) to UE via data radio bearer” to avoid the confusion on UP solution, as tradionally UP solution involves the data transmission between the UE and the UPF.  Meanwhile we have a bit different understanding from Qualcomm on this architecture. Option 1 may present no need to implement AF/UPF function into the RAN entity, since it is also possible to make RAN as a data source for AIML model. Then the only thing left is to specify a DRB to carry such data and send the data to the UE. |
| Lenovo | We see two possibilities to support option 1   * Option 1-a: gNB transfers/delivers AI model to UE via a special DRB (no associated QoS flow) * Option 1-b: gNB first transfers/delivers AI model to CN, and CN transfers/delivers the AI model to UE via UPF/UP (transfer from CN to UE is equivalent to Option 2)   For Option 1-a, RAN2 needs to introduce a new type of DRB, and maybe even a new protocol layer as Xiaomi commented.  For Option 1-b, there will be SA2 impact, and the AI model transfer/deliver between gNB and CN is unclear yet.  We notice some companies also consider the case wherein AF/CN eneities are collocated in gNB in Option 1. However, such case seems identical as Option 2 and Option 4. Maybe better to clarify and probably discuss them separately. |
| Spreadtrum | If option 1-UP solution is considered, the termination of basic data flow is changed from UPF to gNB.  Although gNB can establish DRB and use it for model transfer/delivery. Howerver, there is no QoS flow and no QoS requirement, how UE receives the AIML model and how to guarantee the latency and realibility may need the enhancement to protocol stack, which is complex and needs more discussion. |
| CMCC | We share similar view with vivo and Xiaomi that existing user plane is not applicable since it terminates at UE and UPF. RAN2 needs to discuss how it works first. |
| Nokia, Nokia Shanghai Bell | We agree with Apple, Vivo, Xiaomi, LGE that the existing solution does not support Option 1. |
| ZTE | According to the current scope of the UP protocol, the UP based solution is not crystal clear since the current UP data flow terminated between UPF and UE but not UE and gNB.  But in the study stage, it is so early to say no to each candidate including UP solution, the target for the further discussion is to try to find a UP solution that support model transfer between UE and gNB within the scope of the current protocol stack as much as possible. |
| Ericsson | Echoing what has been commented by other companies. We are also a bit puzzled about how this would work with current UP functioning.  Given the clear CP limitations and, now observing that UP solutions might not work either, we wonder whether RAN2 should start by focusing on Option 4? (i.e., see our input to Q17). |
| Fujitsu | We support the views that UPF is not applicable as the termination for UP data used for model transfer under this option, however, if new function blocks or new layers need to be created just for AI/ML model UP transfer the this option may need to be deprioritized. |
| CATT | Since DRB can not terminate at gNB in current spec, to support UP option1 it may have SA2 impact. Compared with CP option1, the specification impact is much larger. |
| Futurewei | It is clear that the current protocol stack does not support UP data terminates at gNB; it will terminate at UPF at the CN. Even though AF/CN entities may collocate inside gNB under SBA, in reality we are not sure whether this type of implementation is popular. |
| China Unicom | It’s noted that model transfer is transparent to the gNB in the existing specs. So it’s vital to focus on how find a UP solution that involve with gNB without too much specificiation impacts. |
| Mediatek | Based on the comment received so far, it seems that different companies have different assumptions on where the AI/ML model is terminated. If it is assumed that the AI/ML model is terminated at gNB, it’s for sure that the existing protocol doesn’t support model transfer through UP. If it is assumed that the AI/ML model is terminated at certain NF (or AF) in CN, it should be the same as option 2 UP solution. We tend to assume that the AI/ML model is terminated at gNB in this option, and a new type of bearer may be required. |
| Samsung | Similar view to other companies, the existing UP solution has termination points at the UE and UPF, rather than at the UE and gNB, as the case for Option 1 - UP.  Any enhancement to existing UP solution may require input from SA2. |

For this UP solution, it is suggested to collect pros/cons and other comments.

**Q12: Please provide your comments on Option 1 – UP solution in the table below, such as pros/cons, impacts due to model size/latency, use case specific analysis/comments.**

|  |  |
| --- | --- |
| **Company** | **Comments** |
| vivo | **Cons:**   * New UP protocol stack and PDU session management design. |
| Xiaomi | Pros:   * Able to deliver large size model * Less latency compared with LMF/CN solutions   Cons:   * Introduce new protocol layer |
| Qualcomm | Pros-   * Existing 5G architecture can be optimized to achieve collocated UPF and AF (hosting AI/ML models) at the gNB. * Reduces control plane overhead, * Reduces overhead at gNB for model delivery/transfer * Can handle model delivery/transfer during mobility efficiently * Suitable for any size of model transfer * Suitable for transferring multiple models simultaneously * No need to standardize the ML model format in spec   It alleviates all cons of model delivery using a control plane. Also, provide a framework for an optimized model (for the target device) deployment. Note that an optimized model is highly desired for current delay-sensitive use cases. |
| LGE | Pros   * Agreed with QC.   Cons:   * A new application protocol stack that terminates UE and a network entity may need to be introduced to provide model provision services in standardized manner.   SA2 should be involved to make any progress with this option. |
| Apple | We think RAN2 is not in position to list Pros and Cons when it is not clear how this solution works.  Also we believe RAN2 should not be the WG to trigger this discussion.We suggest Proponents to bring this solution to SA2 directly, and send LS to RAN2 if it can be agreed in SA2. |
| OPPO | Pros   * No CP message segmentation issue during AI model transfer/delivery; * No CP message blocking issue during AI model transfer/delivery; * AI model transfer/delivery time delay is smaller than CP solutions; * AI model size is no longer a big limitation for AI model transfer/delivery;   Cons   * gNB will control the AI model transfer/delivery session setup/release, which is usually controlled by CN in traditional procedure; * SA2 coordination may be involved for this Option.   Extra comments:   * AI model transmission authorization/registration procedure may be needed before model transfer/delivery, this may involve SA2 work and common for all 3GPP-based solutions; |
| Dell Technologies | **Pros**:   * Clearly suitable for any AI/ML model sizes and model formats, without special spec impact. * No CP signaling and message impact * Avoid all CP option issues during handover and poor RF conditions   **Cons**:   * Either require a new UP in case UPF is not collocated with the gNB or be mainly suitable for deployments where gNBs are collocated with UPFs. * SA2 involvement may be needed. |
| Huawei, HiSilicon | We agree with Apple that how this solution works should be discussed first, and then Pros/Cons can be discussed. Since this solution is related to other WGs (e.g. SA2), we think that one way is to contact them for more evaluations.  For model characteristics as we comment in Q6, we could have some assumptions in RAN2, and then the impacts on solutions may be different. For example, different model sizes and transmission requirements may correspond to different QoS requirements, and whether it can work depends on the solution details. |
| NEC | We think the main advantegage of this mechanism is (1) able to handle large amount data based AIML model and (2) able to handle the data interruption for AIML model transfer (e.g., caused by radio link failure).  Meanwhile the cost of this mechanism is to introduce a new data radio bearer transmission framework between UE and gNB. Note that the difference from the legacy data service is that this data transmission framework can not be governed by the traditional QoS service model. |
| Lenovo | Option 1-a: gNB transfers/delivers the AI model to UE via a special DRB  Pros:   * AI model size is not limited by RRC message size when comparing to the CP RRC/SRB solution. * Less RRC spec impact   Cons:   * A new type of DRB, and maybe even a new protocol layer managing the AI model is needed. * The AI model transfer/deliver has more delay and is less robust compared to the CP RRC/SRB solution * There is still a need to specify the AI model (at least some parameters)   Option 1-b: gNB first transfers/delivers AI model to CN, and CN transfers/delivers the AI model to UE via UPF/UP (transfer from CN to UE is equivalent to Option 2)  Pros:   * AI model size is not limited by RRC message size when comparing to the CP RRC/SRB solution. * Less RAN side impact * Legacy DRB/PDU session can be used to transfer the AI model   Cons:   * SA2 impacts are foreseen * The AI model transfer/deliver between gNB and CN is unclear yet * The AI model transfer/deliver has more delay and is less robust compared to the CP RRC/SRB solution |
| Spreadtrum | Pros:   * Can transfer/deliver AIML model with large size; * Avoid issues in the option 1-CP solution, e.g., RRC segmentation and continuity in mobility scenario.   Cons:   * Unclear how it works, e.g., latency and realibity requirement. Cross WG spec impact may be needed. |
| CMCC | Pros:  - Support large size model transfer/deliver  - Reduce control plane signaling overhead, no additional RRC issues  Cons:  - A new protocol layer may be introduced  - SA2 may be involved |
| Nokia, Nokia Shanghai Bell | Pros:  - Agree with QC for the pros.  - No additional security or verification required by UE.  Cons:  - Requires new IEs in the RRC to configure UE to start download of a model.  - SA2 input may be needed. |
| ZTE | Pros   * The whole mechanism to guarantee the model transfer with differnet QoS requriement. * No any requirement for the transferred model size, have more adpabilities than CP solution   Consdering there is no clear UP solution on the table so far, the Cons present here is not for sure:  Cons:   * More potential impact on specification than CP solution * For other aspects, it is not clear yet, and depends on the actual UP solution. |
| Ericsson | Agree with Apple. |
| Fujitsu | Pros:   * Can support large-size model transfer/delivery.   Cons:   * How to mitigate the role of UPF is unknown, so many new designs are necessary. |
| CATT | Pros:   * Less impact to RRC spec;   Cons:   * Introduce new UP protocol stack between UE and gNB * RRC layer can not comprehend the model content, and the gNB can not perform delta-configuration to the AI/ML model. |
| Futurewei | Pros:   * Able to handle large-size model without issues * Much less CP overhead, * No or less issue of CP signaling brokage,   Cons:   * Current architecture does not support it which means big standard impact * Need the involvement of SA2. |
| China Unicom | Pros:   * Support large size model transfer/delivery.   Cons:  SA2 involvement may be needed. |
| Mediatek | Pros:   * Support model transfer with large model size     Cons   * Need to support new type of DRB terminiated at the gNB side * Unable to support delta configuration. UE needs to download the whole model when the model/model parameter is updated. * No clear how to initiate and control model transfer process. A mechanism to establish the new type of DRB may be required to enable model transfer. * Not compatible with current mobility procedure. Supporting model transfer during. mobility is not so straightforward. * Require SA2 involvement and evaluation. |
| Samsung | **Pros:**   * Possibility to transfer / deliver large size AI/ML models.   **Cons:**  Require enhancement of existing UP protocol (termination point at gNB). This issue may need to be decided by (or discussed with) SA2. |

#### 2.2.3.2 Option 2 – UP solution

For this UP solution, the principle is that CN (except LMF) can transfer/deliver AI/ML model(s) to UE via UP data. As discussed in phase 1, the applicable use cases are CSI feedback enhancement and Beam management, and there are no specific considerations for Positioning accuracy enhancement.

The basic flow for this UP solution is shown in figure 4 below. For the step “PDU session/DRB establishment”, it may involve the signalling procedures between UE and CN, UE and gNB, and one example is the PDU Session Establishment shown in section A.1 in TS 38.300.



Figure 4: Basic flow for Option 2 – UP solution

**Q13: Regarding Option 2 – UP solution, do companies agree with the principle and the basic flow described above?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
| vivo | Yes |  |
| Xiaomi | Yes |  |
| Qualcomm | Yes |  |
| LGE | Yes |  |
| Apple | Yes |  |
| OPPO | Yes |  |
| Dell Technologies | Yes |  |
| Huawei, HiSilicon | Yes |  |
| NEC | Yes |  |
| Lenovo | Yes with comment | For Option 2, we also see two possible alternatives   * Option 2-a: The AI model is transferred/delivered by legacy DRB/PDU session, and is received/handled by UE application layer, or * Option 2-b: The AI model is transferred/delivered by a special DRB/PDU session, and is received/handled by UE access stratum   We believe both can be considered with different implication at this stage. |
| Spreadtrum | Yes |  |
| CMCC | Yes |  |
| Nokia, Nokia Shanghai Bell | Yes |  |
| ZTE | Yes |  |
| Ericsson | Yes |  |
| Fujitsu | Yes |  |
| CATT | Yes |  |
| Futurewei | Yes | This can be the baseline. |
| China Unicom | Yes |  |
| Mediatek | Yes |  |
| Samsung | Yes |  |

For this CP solution, it is suggested to collect pros/cons and other comments.

**Q14: Please provide your comments on Option 2 – UP solution in the table below, such as pros/cons, impacts due to model size/latency, use case specific analysis/comments.**

|  |  |
| --- | --- |
| **Company** | **Comments** |
| vivo | **Pros:**   * Limited or no RAN2 impact for model transfer.   **Cons:**   * The latency of model transfer and switching during handover may not be guaranteed, especially when the model is valid per cell. * SA2 needs to study which CN functionality is responsible for the model transfer and how to set up the connection between CN functionality and UE. |
| Xiaomi | Pros:   * May be able to deliver large size model, up to SA2 evaluation. * May require less signalling exchange during handover, if AI model is applicable in multiple cells   Cons:   * Large latency compared with gNB solution. * Require more inter-layer signalling compared with gNB solution, since the use case is mainly located in physical. * Not clear whether CN node is able to determine the applicable AI for physical use case   However, it’s better leave to SA2 to evaluate the feasibility and Pros/Cons. We suggest RAN2 to focus on other solution. |
| Qualcomm | Pros:   * Alleviates all cons of model delivery/transfer associated with model delivery using NAS signaling * Avoids the need to standardize the ML model format in spec. * The network can provide different 5Qis for ML model transmission if it has different QoS requirements. * 5G architecture provides flexibility to place UPF at any desired network entity. This help in avoiding duplication of models at multiple places. One can achieve the right balance between the required model delivery/transfer in latency and model storage in the network. |
| LGE | Pros   * Agreed with QC.   Cons:   * A new application protocol stack that terminates UE and a network entity may need to be introduced to provide model provision services in standardized manner. * SA2 should be involved to make any progress with this option. |
| Apple | **Pros:**  - Limited or minor RAN2 spec impact  - Native support of large size AI/ML model transfer.  **Cons:**  - The signaling latency is higher than CP option 1 (i.e. gNB solution)  - It may have multi-vendor interoperation issues  - DRB transmission is generally less robust than SRB (assuming gNB is not aware of AI/ML model transfer in one DRB as in legacy).   * Some SA2 enhancement may allevate this cons, e.g. a new 5QI for AI/ML model. But it is out of RAN2 scope.   **Impacts to other WGs:**  - SA2 and CT1, including   * New 5QI for AI/ML model. * Which 5GC entity (e.g. AMF, or SMF, or LMF, or a new entity) to communicate with UPF to establish PDU session for AI/ML model transfer.   **Suitable use case:**  - Transfer of offline training model for all use cases.  - Transfer of inference model for AI/ML based Positioning  **Question for some above comment:**  - Why this solution can alleviate standard efforts on model format in spec? Does it mean only existing indurtial AI/ML mode formats (e.g. .h5) can be used? |
| OPPO | Pros   * The existing UP resource setup/release framework can be reused; * Less AS spec impact compared to Option 1 – UP solution; * No CP message segmentation issue during AI model transfer/delivery; * No CP message blocking issue during AI model transfer/delivery; * AI model size is no longer a big limitation for AI model transfer/delivery.   Cons   * Longer time delay compared to Option 1 – UP solution;   Extra comments:   * AI model transmission authorization/registration procedure may be needed before model transfer/delivery, this may involve SA2 work and common for all 3GPP-based solutions; |
| Dell Technologies | Pros:   * Same as QC, LGE * No to little RAN2 impact   **Cons**:   * Less robust AI/ML model delivery/transfer on a DRB than SRB. Thus, AI/ML model delivery/transfer reliability will be under question. * Coordination with SA2 is needed. |
| Huawei, HiSilicon | Same as our response to Q12. |
| NEC | If 5GC or a server somewhere at data domain holds the AIML model, UP based transmission should be the only way, considering that this is the traditional approach to transmit the data between the UE and the network.  The advantage of this model is to to simply reuse the existing user plane data transfer framework, removing the need to introduce the signalling support.  The disadvantage of this model is that the applicablitiy is highly depending on the fact that the 5GC domain should holds the AIML model, which means it may apply to all of the gNB/cell in the network. However, in practice, different gNB may hold different AIML model for the same use case. |
| Lenovo | Option 2-a: The AI model is transferred/delivered by legacy DRB/PDU session, and is received/handled by UE application layer  Pros:   * AI model size is not limited by RRC message size when comparing to the CP RRC/SRB solution. * Less RRC spec impact   Cons:   * The AI model transfer/deliver has more delay and is less robust compared to the CP RRC/SRB solution * Depending on which entity trains the AI model (and provides to CN), there are different implications:   + If the AI model is trained by the gNB and later provides to CN, the AI model transfer between gNB and CN is unclear yet   + If the AI model is trained by CN, how does a CN node obtain all necessary training data (e.g., L1/L3 RAN measurements) is tricky. In legacy, the exposure of RAN measurements to CN is quite limited. Also, how to get gNB involved is also tricky.   Option 2-b: The AI model is transferred/delivered by a special DRB/PDU session, and is received/handled by UE access stratum  Pros:   * AI model size is not limited by RRC message size when comparing to the CP RRC/SRB solution. * Less RRC spec impact   Cons:   * SA2 impacts are foreseen * A new type of DRB, and maybe even a new protocol layer managing the AI model is needed. * The AI model transfer/deliver has more delay and is less robust compared to the CP RRC/SRB solution * There is still a need to specify the AI model (at least some parameters) |
| Spreadtrum | Pros:   * Can transfer/deliver AIML model with large size; * Avoid issues in the option 1-CP solution, e.g., RRC segmentation and continuity in mobility scenario.   Cons:   * It is unclear which CN entities can store RAN AIML models, i.e., CSI compression and prediction, beam management. SA2 may need to discuss. * Larger latency than option 1-CP solution. |
| CMCC | Pros:  - Support large size model transfer/deliver  - No/limited spec impacts on RAN2  Cons:  - Same concern as Q8, we are not sure that whether it is suitable for CN to determine AI/ML model for physical use case, it requires more SA2 evaluation  - Higher latency compared to option 1 |
| Nokia, Nokia Shanghai Bell | Pros:  - Highly desirable for large size ML models.  - This is possible because CN can perform PDU session establishment which enables to transfer large size model using DRBs.  Cons:  - CP signalling is needed to configure and initiate the model transfer from the CN.  - Option 2 relates to models for the gNB-centric use cases of beam management and CSI compression, which are out of scope for CN.  - Coordination would be required between the gNB and CN function (to be determined) such that a gNB-specific model or a model specific to a group of gNBs would be properly selected. The gNB would also need to be notified of the model transfer/delivery so that the gNB could activate, deactivate, and switch between models. How this coordination would work has not been discussed. |
| ZTE | Pros   * The whole mechanism to guarantee the model transfer with differnet QoS requriement. * No any requirement for the transferred model size, have more adpabilities than CP solution * Less RAN specifcation impact   Cons   * The most work may need to be done in SA2 , for example, UPF funtion extents or/and add an OAM assoaciated with the UPF. |
| Ericsson | We understand the pros listed by Qualcomm. But wonder on the concrete RAN2 involvement in such a solution. On this matter, and as highlighted before, we see no real need or benefit to bring into this discussion other WGs without TUs allocated for this SI.  For which we propose RAN2 to analyse other solutions. Alternatively, as mentioned by other companies, proponents could start by trigging such discussion on those other concerning WGs first. |
| Fujitsu | Pros:   * Similar to Q12 * Less RAN2 specification impact   Cons:   * The latency issue. * Need to consider the mitigation of the LCM roles played by CN and gNB. |
| CATT | Pros:   * Less RRC impact, no complex operation such as segmentation needs to be considered in RRC.   Cons:   * The latency may be increased compared with UP based solution option 1. |
| Futurewei | Pros:   * Able to transfer/deliver models of any size; * Limited spec impact   Cons:   * Longer delay comparing to CP-based solutions and UP/gNB-based solution; * Multi-vendor interoperability may be an issue as the gNB has no control over the models used, as well as LCM; |
| China Unicom | Pros:   * Similar to Q12 * Less RAN2 impact.   Cons:  • Not feasible for real-time AI service.  • SA2 involvement may be needed. |
| Mediatek | Pros:   * Support model transfer with large model size * Less RAN2 specification impact     Cons   * Longer latency of model transfer compared with option 1 * Unable to support delta configuration. UE needs to download the whole model when the model/model parameter is updated. * No clear how to initiate and control model transfer process. * Not compatible with current mobility procedure. Supporting model transfer during mobility is not so straightforward. * Require SA2 involvement and evaluation.   We tend to agree that model transfer from CN may not be proper for the use cases of AI/ML operation purely over air interface, e.g. for CSI and BM, requiring RAN to be responsible for the life cycle management. How to make RAN node be aware of AI/ML model needs to be considered further. |
| Samsung | **Pros:**   * Possibility to transfer / deliver large size AI/ML models. * Limited impact to RAN2 for model transfer/delivery.   **Cons:**   * Large latency compared with Option 1 – UP (i.e. UP termination at gNB solution).   Potential impact to other WGs, for example, SA2 and CT1. |

#### 2.2.3.3 Option 3 – UP solution

As discussed in phase 1, for this UP solution, the applicable use case is Positioning accuracy enhancement.

For this UP solution, during phase 1 discussion, one company pointed out one UP solution according to TR 23700-71, and some companies pointed out that the solution details are not clear, so that more discussions are needed. So it is suggested to collect companies’ views on the principle and the basic flow.

**Q15: Regarding Option 3 – UP solution, please companies provide your views on the principle and the basic flow in the table below.**

|  |  |
| --- | --- |
| **Company** | **Comments** |
| vivo | Similar to Figure 4 in Option 2, the UE will establish the PDU session with LMF via UPF and the only difference is the CN entity is LMF. |
| Xiaomi | We understand the UP solution in TR 23700-71 is the LPP can be delivered in the UP, so the UP solution for delivering/transferring the AI model is the same as the option 3-CP solution. |
| Qualcomm | Option 3 is a sub-case of option 2. |
| Apple | We suggest to merge this solution with option 2.  Meanwhile, please note that this solution is only included in SA2 technique report (TR 23700-71). Whether it becomes SA2 Rel-18/19 normative work should be decided in SA plenary. We believe RAN2 should not make conclusion / assumption on this solution. |
| OPPO | We agree that this option may be a special case for Option2-UP solution,we can first focus on Option2-UP solution. |
| Dell Technologies | We suggested merging Option 2 with Option 3. |
| Huawei, HiSilicon | Firstly, we think this solution may be similar to Option 2 – UP solution, and we agree with Apple that whether the UP solution(s) in TR 23700-71 will become SA2 normative work needs more discussions in other WGs.  Secondly, same as our response to Q12, we think that how this solution works should be discussed first, and then Pros/Cons can be discussed. |
| NEC | We agree with other companies on that there is no need to take specific discussion for this approach, since it may be covered by option 2. |
| Lenovo | If the UP solution for LPP is supported by SA2, we understand it will be eventually similar as Option 2, while the CN entity will be LMF in this case. |
| Spreadtrum | Agree with above companies that this solution is similar to Option 2 – UP solution. |
| CMCC | We share similar view with Apple. |
| Nokia, Nokia Shanghai Bell | We cannot properly discuss the UP solution for the LMF until it is decided by SA2. It is also unclear how the model delivery would be triggered with a solely UP solution. |
| ZTE | Share the same view with Qualcomm and apple. |
| Ericsson | Agree with other companies, this Option should be a subcase of Option 2. |
| Fujitsu | Agree with Apple, we do not think it is a better idea to discuss this option separately until SA2 has more progress. |
| CATT | We agree with other companies that the Option 3 can be a sub-option of Option2, since both options should transmit the AI model via UPF. |
| Futurewei | This is similar to Option 2. |
| China Unicom | It’s a sub-option of Option 2. |
| Mediatek | Option 3 is a sub-case of option 2. |
| Samsung | We agree that Option 3 –UP could be merged with Option 2 - UP. |

For this UP solution, it is suggested to collect pros/cons and other comments.

**Q16: Please provide your comments on Option 3 – UP solution in the table below, such as pros/cons, impacts due to model size/latency, use case specific analysis/comments.**

|  |  |
| --- | --- |
| **Company** | **Comments** |
| vivo | **Pros:**   * Limited or no RAN2 impact for model transfer. * Compared to Option 2 (i.e., UP solution between other CN entities and UE), less SA2 impact is foreseen as option 3 reuses LMF to handle the model transfer.   **Cons:**   * The latency of model transfer and switching during handover may not be guaranteed, especially when the model is valid per cell. |
| Xiaomi | As commented in the Q15, we think the option 3-UP is the same as the option 3- CP, both solutions are to enhance the LPP. |
| Qualcomm | Same as the response to Q15. |
| Apple | We sugges to merge it with option 2. |
| OPPO | The same view as Q15. |
| Dell Technologies | Same as our response to Q15. |
| Huawei, HiSilicon | Same as our response to Q15. |
| NEC | See reply for Q15 |
| Lenovo | Same as for Option 2 |
| Spreadtrum | The same view as Q15 |
| CMCC | Same as the response to Q15. |
| Nokia, Nokia Shanghai Bell | Same response as in Q15 |
| ZTE | Same reponse as in Q15 |
| Ericsson | Same as answer to Q15. |
| Fujitsu | Same response as in Q15. |
| CATT | UP option3 is similar with the CP option3, for which different paths are utilized to transfer the AI/ML model between UE and LMF. Both options have no RRC impact. |
| Futurewei | Similar to Option 2 |
| China Unicom | See comments in Q15. |
| Mediatek | Same reply as for Q14. |
| Samsung | See answer to Q15. |

### 2.2.4 Option 4

For this option, as discussed in phase 1, the applicable use cases are CSI feedback enhancement, Beam management, and Positioning accuracy enhancement.

For this option, it may be transparent to 3GPP and it can be left to implementation. In this case, it seems appropriate to directly collect companies’ views on this solution.

**Q17: Please provide your comments on Option 4 in the table below, such as pros/cons, impacts due to model size/latency, use case specific analysis/comments.**

|  |  |
| --- | --- |
| **Company** | **Comments** |
| vivo | **Pros:**   * No 3GPP impact for model transfer.   **Cons:**   * The latency of model transfer and switching during handover may not be guaranteed, especially when the model is valid per cell. * The model is transparent and out of control for the operators. Different implementations may lead to different model performances and a huge burden of model management (e.g., frequent model activation/deactivation). * Multi-vendor interoperability is almost impracticable, as massive offline coordination is needed, especially for the CSI compression use case. |
| Xiaomi | Cons:   * Large latency compared with gNB solution. * Require more inter-layer signalling compared with 3GPP solution, since the use case is mainly used in 3GPP protocol. * Not clear which node is responsible to determine the applicable AI * Difficult to do the IOT test between UE and NW vendors |
| Qualcomm | Option 4 is the same as option 3 from an architectural viewpoint.  We agree that for model delivery/transfer there may not be the 3GPP impact. However, there are closely correlated LCM aspects that need to be studied. For example, although the model delivery/transfer is transparent to the 3GPP network, model selection may happen at the gNB, i.e., gNB may determine which model UE needs to download and use.  In a model ID-based LCM, in this option the following need to be studied in RAN2:   * UE capability signaling to indicate supported model ID * Configuration by the gNB based on the indicated UE capability * Model delivery/transfer based on configuration, if not available at the UE * Other LCM procedures, e.g., activation deactivation, switching, and fallback   As I previously mentioned, the SBA provides flexibility to deploy the model at any place within the 3GPP network/outside the 3GPP network. Therefore, to achieve a lower latency, the OTT server may be placed close to the RAN. Furthermore, the model control can happen based on the meta info. This we agreed on in the RAN2#119 meeting. Therefore, even if models are placed in the OTT server, the 3GPP network (operator) may still control the usage of the model. |
| LGE | We think the difference between the UP-based solution discussed above and O4 is non-standard in 3GPP not specifying the protocol stack related to ML.  However, O4 also seems to require interaction between UE sublayers (e.g., RRC) and non-standard stacks. |
| Apple | First, we agree with QC that Option 4 is the same as UP option 2 from an architectural viewpoint, and 3GPP edge computing solution can be reused. But Option 4 still has below differences:  1) For UP Option 2, CN may be aware of AI/ML model. For example, a new 5GC entity to manage AI/ML models. But Option 4 can't.  2) For UP Option 2, gNB may also be aware of AI/ML model transfer in one DRB. For example, a dedicated 5QI is reserved for AI/ML model transfer. But option 4 can't.  With above clarifcaition, our understanding on Pros and Cons:  **Pros:**  - Limited or minor RAN2 spec impact   * We agree with QC that model ID and LCM may have impacts.   - Native support of large size AI/ML model transfer.  **Cons:**  - The signaling latency is higher than CP option 1 (i.e. gNB solution). Whether it is even higher than UP option 2 (i.e. UPF solution) depends on whether/how it is edge deployed.  - It will have multi-vendor interoperation issues.   * We alsp agree with QC that model ID and meta info are useful to alleviate this issue.   - DRB transmission is generally less robust than SRB   * Different from UP option 2, CN and gNB are not aware of AI/ML model being transfered in one DRB. Thus, no extra robust protection.   **Impacts to other WGs:**  - None  **Suitable use case:**  - Transfer of offline training model for all use cases. |
| OPPO | Pros:   * No spec impact is involved for model transfer/delivery; * Model size is no longer a big limitation.   Cons:   * Multi-vendor issue should be considered as the network has no idea about which AI model is deployed at UE side if UE acquires AI model from the server based on implementation; * Model transfer/delivery latency can not be guaranted for this Option. |
| Huawei, HiSilicon | Firstly, we think this option is transparent to 3GPP and it can be left to implementation.  Secondly, for inter-operability impacts, the solutions mentioned in section 2.2.2 and 2.2.3 have no problems, but this Option 4 is quite challenging.  Secondly, since it is an implementation related solution, we do not think it can work well for some evaluation metrics (as we comment in Q6). For example, for model size, there may be large model and small model; for transmission latency, there may be low/high latency requirements, how the server co-ordinates with 3GPP networks in order to meet these requirements is not clear. |
| NEC | Even though the model transfer itself may be transparent to 3GPP, however the underlined 3GPP data tunnel needs to be established in order to enable such model transfer. Then we suggest also to identify the possible standards impact for this option. |
| Lenovo | We have similar understanding as Apple w.r.t the difference between Option 2 and Option 4. If we assume 3GPP NW is aware of the AI model transfer/delivery in Option 4, we may eventually talking about Option 2.  Pros:   * Minimal spec impact (it could work without 3GPP network involvement) * AI model is not limited by any size   Cons:   * The model transfer/delivery is transparent to gNB/CN, it could be tricky to get gNB/CN involved in the AI model LCM. It could be problematic when the network needs to be in control of what happening at the UE side and especially in two-sided models where one side of the model is intended to be located at the network side. * The AI model transfer/deliver has more delay and is less robust compared to the CP RRC/SRB solution |
| Spreadtrum | Pros:   * Little RAN2 impact. * Can transfer/deliver AIML model with large size;   Cons:   * Generally have large latency compared with 3GPP solution. * Unclear how to perform model LCM, e.g., model update. * As for inter-layer signalling, UE should decode the received packets and deliver them to lower layers, finally those packets should be used to recovery the AIML model who works in PHY. It seems that inter-layer signalling should also be considered in other options. |
| CMCC | Pros:  - No 3GPP impacts for model transfer/delivery  - Support large size model  Cons:  - There is multi-vendor interoperability issue.  - The latency may not be guaranteed and depends on which entity performs model transfer/delivery.  - It is difficult for network to perform model control, such as model activation/deactivation/fallback, because the network has no idea on which model is deployed at UE. |
| Nokia, Nokia Shanghai Bell | Pros:  - Actual model delivery/transfer is transparent to 3GPP signalling.  - There is no limitation to model size, or the number of models transferred  Cons:  - Because the transfer/delivery is not controlled by the network, LCM-based functionality signalling will be required to facilitate interoperability between the UE and the network as well as model identification and verification.  - When network cannot control the model transfer/delivery, the transfer of large model may impact important and delay sensitive user data traffic.  - How to synchronize network and server, so that network can take needed actions, requires further studies and is not fully under 3GPP control. |
| ZTE | Pros:   * 3GPP transparent, no any specifciation impact. * No need to spend the time to discuss * The Highest freedom degree of UE implemention   Cons:   * model transfer with this option is out of control with the NW. * Multi-vendor interoperability issue. |
| Ericsson | Firstly, we think it is important to clarify what Option 4 represents. To us, Option 4 is simply letting an eventual model transfer/delivery up to implementation, i.e., without 3GPP-spec impact.  Note that model transfer/delivery is still under discussion in RAN1. And, here, we understand that not just the requirements of such a procedure are under analysis, but the actual benefits or feasibility of it (e.g., for the CSI use case). On this, we can echo the views of other companies, that a solution according to Option 4 (i.e., up to implementation) could represent an interoperability issue. But this does not mean that the solution is unfeasible.  On the contrary, and taking on what Qualcomm have written above, RAN2 could analyse the related (LCM) procedures where an “over-the-top” solution is manageable or unfeasible according to what the SI is set to accomplish.   Under that reasoning, and acknowledging the challenges listed for all other alternatives, we believe that RAN2 could actually start its study by addressing Option 4. |
| Fujitsu | Pros:   * Save all the signaling and procedure overhead for model transfer/delivery. * Vendors can freely manipulate the details of models (including model size) without NW constraints.   Cons:   * It has to be considered with other LCM functions, if option 4 is chosen then many other LCM functions (e.g., model identification) has to follow this way for alignment, in one case, the LCM system has to be re-designed or even cancelled. * In this case NW can do nothing expect for data collection. * Multi-vendor interoperability issue. |
| Futurewei | Pros:   * No or minimal spec impact; * Support of large-size models; * Flexibility of implementation;   Cons:   * May not be able to avoid spec impact entirely; * Most inter-layer signaling needed for models to be used in PHY; * Longer latency and hard to guarantee the delivery/transfer in time; * The LCM of models and even the quality of the model will be out of control of the network/operator; * Without knowing what model is used, multi-vendor issue will be challenging, or requires lots of coordinations among vendors. |
| China Unicom | Pros:   * No 3GPP spec impacts.   Cons:   * Can’t be controlled by the network. |
| Mediatek | We share the same view as Lenovo that if the model transfer is totally transparent to 3GPP, one basic question is why we need to involve network, what kind of benefits can be obtained with LCM controlled by network and how it works?  I am confused by Qualcomm’s statement that Option 4 is the same as option 3 from an architectural viewpoint. If the system architecture for Option 4 is the same as option 3, model transfer is actually not transparent to CN and SA2 needs to be involved. SA2 may need to consider the cases that the OTT server is a trusted server or an untrusted server.  Pros:   * Flexible to support various side of model transfer/deliver   Cons:   * Unable to support delta configuration. UE needs to download the whole model when the model/model parameter is updated. * Not compatible with current mobility procedure. Supporting model transfer during mobility is not so straightforward. * The need of network involvement for LCM is not clear. * It may still require SA2 involvment and evaluation |

# 3 Conclusion

In phase 1, the summary proposals are listed as below:

**Proposal 1: Use the wording “model transfer/delivery” for the RAN2 study.**

**Proposal 2: Agree to discuss the following solutions in phase 2:**

* **Option 1: Model transfer/delivery between UE and gNB via CP and UP solutions**
* **Option 2: Model transfer/delivery between UE and CN (except LMF) via CP and UP solutions**
* **Option 3: Model transfer/delivery between UE and LMF via CP and UP solutions**
* **Option 4: Model transfer/delivery between UE and server**

**Proposal 3: The discussion on model delivery between network entities is postponed.**

**Proposal 4: Agree on Table 2a for the RAN2 study and it can be used for further discussions.**

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

|  |  |
| --- | --- |
| **Architectural assumptions** | **Applicable use cases** |
| Option 1 | CSI feedback enhancement  Beam management  Note: No specific considerations for Positioning accuracy enhancement for this option. |
| Option 2 | CSI feedback enhancement  Beam management  Note: No specific considerations for Positioning accuracy enhancement for this option. |
| Option 3 | Positioning accuracy enhancement |
| Option 4 | CSI feedback enhancement  Beam management  Positioning accuracy enhancement |

[Phase 2 summary will be added later]

# 4 References

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