**3GPP T****SG-RAN WG2 Meeting #123 R2-230xxxx**

**Toulouse, France, 21- 25August, 2023**

**Agenda item: 7.16.2.1**

**Source: CMCC**

**Title: Report of [Post122][060][AIML] Mapping of functions to physical entities (CMCC)**

**Document for: Discussion and Decision**

# 1 Introduction

This contribution is aimed at reporting the discussion and results of the following post email discussion:

* [Post122][060][AIML] Mapping of functions to physical entities (CMCC)

Scope: Starting from relevant contents in R2-2305613, attempt to produce an agreeable description of Mapping of functions to physical entities. UP to rapporteur to structure

Intended outcome: Report

Deadline: Long (4th Aug, 10:00 UTC)

For this email discussion, the outcome is expected to be used for discussions of possible solutions and specification impacts. The listed entities do not mean that any specification impacts for the involving entities.

As indicated by the Chair, the inactive period is:

July 1st – 30th 3GPP Inactive Period

The deadline is 4th Aug, 10:00 UTC. The summary will be provided by 8th Aug 10:00 UTC, and then companies can check it. The final summary will be submitted by the submission deadline (11th Aug).

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

|  |  |  |
| --- | --- | --- |
| **Company** | **Name** | **Email Address** |
| Apple | Peng Cheng | pcheng24@apple.com |
| OPPO | Jiangsheng Fan | fanjiangsheng@oppo.com |
| Mavenir | Fan Yang | Fan.yang@mavenir.com |
| vivo | Boubacar Kimba | kimba@vivo.com |
| Xiaomi | Xing Yang | Yangxing1@xiaomi.com |
| Mediatek | Yuanyuan Zhang | Yuany.zhang@mediatek.com |
| Lenovo | Congchi Zhang | Zhangcc16@lenovo.com |
| CATT | Da Wang | wangda@catt.cn |
| Qualcomm | Rajeev Kumar | rkum@qti.qualcomm.com |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# 2 Discussion

In last meeting, RAN2 agreed a general architecture covering the model based and/or functionality based LCM as shown in Fig.1. The general AI/ML framework consists of model training, inference, management, data collection and model storage.



Fig.1 General architecture for AI/ML

For model storage, RAN2#122 agreed that Model Storage in the figure is only intended as a reference point (if any) for protocol terminations etc for model transfer/delivery etc, and it is not intended to limit where models are actually stored. For simplicity, it is assumed that the entity of model training and model storage is the same one. Thus, mapping of model storage to entities is not discussed in this email discussion.

For management, RAN2 agreed it may be model based management or functionality based management in last meeting. Based on previous discussion and RAN1’s progress, the rapporteur understands that management includes monitoring, selection, (de)activation, switching, fallback. In RAN1#113 meeting, RAN1 agreed that for model ID-based and functionality based LCM, the same or similar procedures may be used for their activation, deactivation, switching, fallback, and monitoring. To make it clear during the discussion, the rapporteur suggests to use model/functionality monitoring and control (selection, (de)activation, switching, fallback) instead of management.

|  |
| --- |
| **RAN1#113 Agreement**  For functionality/model-ID based LCM,   * Once functionalities/models are identified, the same or similar procedures may be used for their activation, deactivation, switching, fallback, and monitoring. |

Considering data collection is discussed in sub-agenda 7.16.2.2, it is suggested to focus on the following LCM purposes in this email discussion:

* Model training
* Inference
* Model transfer/delivery
* Model/functionality monitoring
* Model/functionality control, including selection, (de)activation, switching, fallback

For different use case with different model type (e.g. UE-side model, two-sided model, gNB-side model, LMF-side model), the mapping of functions to physical entities may be different, so it is better to discuss per use case per model type.

In this email discussion, the rapporteur suggests to focus on non-split gNB architecture in this stage to make it clear and simple, i.e. CU-DU architecture is not considered in this email discussion.

## 2.1 CSI feedback enhancement

For CSI feedback enhancement, RAN1 agreed to study the following sub-use cases:

• Spatial-frequency domain CSI compression using two-sided AI model

• CSI prediction using UE-side model

### 2.1.1 CSI compression with two-sided model

* Model training and model transfer/delivery:

For CSI compression using two-sided AI/ML model use case, RAN1 agreed to further study AI/ML model training collaboration Type 1 and Type 3, and Type 2 is de-prioritized in R18 SI.

|  |
| --- |
| **RAN1#110 Agreement**  In CSI compression using two-sided model use case, the following AI/ML model training collaborations will be further studied:   * Type 1: Joint training of the two-sided model at a single side/entity, e.g., UE-sided or Network-sided. * Type 2: Joint training of the two-sided model at network side and UE side, repectively. * Type 3: Separate training at network side and UE side, where the UE-side CSI generation part and the network-side CSI reconstruction part are trained by UE side and network side, respectively.   **RAN1#111 Conclusion**  In CSI compression using two-sided model use case, training collaboration type 2 over the air interface for model training (not including model update) is deprioritized in R18 SI. |

For Type 1, the two-sided AI/ML model which includes a CSI generation part and a CSI reconstruction part is trained at network, and then the trained UE-side CSI generation part will be transferred from network to the UE, or vice versa. The model transfer/delivery between UE and network via air interface is needed for Type 1.

For Type 3, the UE-side CSI generation part and the network-side reconstruction part are trained by the UE side (may includes UE and OTT server) and network side separately. For example, the two-sided AI/ML model is trained at network, then the network sends the data set which includes input and output to the UE, and the UE side trains a UE-side CSI generation part using the data set. On the other hand, it is possible that the UE side trains the two-sided model and sends the data set to the network, and network side trains the network-side reconstruction part based the data set.

At RAN2#122, RAN2 assumed that for model training, training data can be generated by UE/gNB and terminated at gNB/OAM/OTT server. It is suggested that model training can reside at gNB/OAM/OTT server. Therefore, for training type 1, the model transfer/delivery can be from gNB to UE, or from OAM to gNB and UE, or from OTT server to gNB and UE. And for training type 3, the UE-side model can be transferred/delivered from OTT server to UE if it is trained at OTT server; for NW-side model, there is no model transfer/delivery if the NW-side model is trained at gNB, or the NW-side model is transferred/delivered from OAM to gNB if it is trained at OAM.

* Model inference:

For two-sided AI/ML model, it is obvious that model inference reside at UE and gNB side for UE-side CSI generation part and NW-side CSI reconstruction part separately.

* Model/functionality monitoring and control:

RAN1 agreed that NW and UE can both monitor the performance, NW makes the decisions of model control.

|  |
| --- |
| **RAN1#110b Agreement**   * NW-side performance monitoring: NW monitors the performance and make decisions of model activation/ deactivation/updating/switching * UE-side performance monitoring: UE monitors the performance and reports to Network, NW makes decisions of model activation/ deactivation/updating/switching |

Based on the above analysis and RAN1 agreements, the possible mapping of functions to physical entities for CSI compression with two-sided model is shown in following table.

Table 2.1-1: The mapping of functions to physical entities for CSI compression with two-sided model

|  |  |  |
| --- | --- | --- |
|  | **AL/ML functions (if applicable)** | **Mapped entities** |
| a) | Model training | gNB, OAM, OTT server |
| b) | Model transfer/delivery | For training Type 1: gNB->UE, or OAM->gNB&UE, or OTT server->gNB&UE  For training Type 3: For UE-side model, OTT server->UE if the UE-side model is trained at OTT server; For NW-side model, no model transfer/delivery if the NW-side model is trained at gNB, or OAM->gNB if the NW-side model is trained at OAM; |
| c) | Inference | NW-side: gNB  UE-side: UE |
| d) | Model/functionality monitoring | NW-side: gNB  UE-side: UE |
| e) | Model/functionality control (selection, (de)activation, switching, fallback) | gNB |

Note 1: For a), only data collection part may be further discussed.

Note 2: Whether/how OAM is to be involved may need to consult SA5.

Q1: Do you agree the mapping of functions to physical entities for CSI compression with two-sided model in Table 2.1-1?

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Yes for which bullet(s)** | **No for which bullet(s)** | **Comments** |
| Apple | a), b), c), d) (but b/c/d has terminology issue) | e)  And dataset transfer is missed | For b)/c)/d), one terminology issue: since two-sided model is used for CSI compression, the term UE side/NW side model is not accurate:   * "UE side model" -->"UE part of two-sided model" * "NW side mode" --> "NW part of two-sided model"   For e), UE autonomous model switch was agreed in RAN1#112, as highlighted below:   * In functionality-based LCM   + Network indicates activation/deactivation/fallback/switching of AI/ML functionality via 3GPP signaling (e.g., RRC, MAC-CE, DCI).   + Models may not be identified at the Network, and UE may perform model-level LCM.     - Study whether and how much awareness/interaction NW should have about model-level LCM * In model-ID-based LCM, models are identified at the Network, and Network/UE may activate/deactivate/select/switch individual AI/ML models via model ID.   Meanwhile, during RAN1 discussion on AI 9.2.2.2, another solution is: gNB can configure a set of parameters, and UE will choose the UE part model within the parameters configuration, and let gNB know the corresponding NW-part model to use.  Thus, We don't see any reason that RAN2 can preclude one RAN1 agreed direction. So, e) should be updated to " gNB, UE".  Meanwhile, we think dataset transfer is missed in the table. Note that it doesn't mean how NW to store dataset but it means which NW entity needs to receive dataset for training/inference/monitoring purpose. It may have spec impacts. Thus, we suggest to add row f)  f): dataset transfer:   * For training Type 3: For UE-side model, from UE to OTT server (if model is trained in OTT server), and from gNB to UE; For NW-side mode, from UE to gNB (if trained in gNB), or from UE to OAM (if trained in OAM) * For Monitoring: if NW monitors UE-side model, from UE to gNB; if UE monitors NW-side mode, from gNB to UE. |
| OPPO | All with comments and modification |  | For bullet a, we should make it clear that only offline model training is focused in this email discussion, this principle should be applied to all subsequent questions. Another point is that even if RAN2 agreed that training data can be generated by UE/gNB and terminated at gNB/OAM/OTT server in last RAN2 meeting, it still does not preclude the case that model is trained at UE side, the above RAN2 agreement only emphasize the possible training data exchange with other nodes, UE still can collect the training data and use it for model training without training data exposed to other node based on the following RAN1 agreements.    For bullet b, as clarified for bullet a, it’s still possible for UE to train the model.  so for training Type 1, the missing part is that UE->gNB, which is not explicitly precluded by RAN1 so far, we should consider this scenario also;  For training Type 3, for UE side model, the missing part is that no model transfer/delivery if the UE-side model is trained at UE, For NW-side model, the possible missing part is that OTT->gNB if the NW-side model is trained at OTT, to differentiate with training Type 1 case, i.e. OTT server->gNB&UE, for training Type 3, network side OTT server is different than the OTT server to train UE side model, people should consider whether to consider this scenario also for training Type 3 as for training Type 1 we also consider gNB side model is trained at OTT server.  For bullet d, firstly we’d like to clarify this is only for real-time model monitoring as SA5 is also considering something for non-real time model monitoring, which may or may not have RAN2 impact, to be safe, we may need to clarify this for all subsequent questions;  Another point is that for UE side model monitoring, based on above RAN1 agreement, even if UE-side model monitoring happens at UE side, UE still needs to report the collected metrics to network side, to make it clear, we have the following suggestion:  NW-side: gNB, i.e. gNB directly collects the model monitoring metrics.  UE-side: UE, i.e. UE directly collects full or partial model monitoring metrics and reports the collected metrics to gNB side. |
| Mavenir | Yes for (c)(d) but comments and modification for (a)(b) (e) |  | For a), Training Type conclusion was made in RAN1#110, as below:  In CSI compression using two-sided model use case, the following AI/ML model training collaborations will be further studied:  • Type 1: Joint training of the two-sided model at a single side/entity, e.g., UE-sided or Network-sided.  • Type 2: Joint training of the two-sided model at network side and UE side, repectively.  • Type 3: Separate training at network side and UE side, where the UE-side CSI generation part and the network-side CSI reconstruction part are trained by UE side and network side, respectively.  Since model can be trained at UE side, we suggest a) should be updated to “gNB, OAM, OTT server,UE”, and b) should be updated to “For training Type 1: gNB->UE, or OAM->gNB&UE, or OTT server->gNB&UE, UE->gNB”  Model/functionality control (selection, (de)activation, switching, fallback) can be implement at UE, we suggest e) should be updated to “gNB,UE”  For Apple’ s proposal of f) ,we suggest below   * For type1 training, For NW model, dataset is from UE to gNB(if model is trained at gNB), dataset is from UE to OAM(if model is trained at OAM),dataset is from UE to OTT server(if model is trained at OTT server) * For type3 we agree with Apple’s opinion |
| vivo | All with comments on a)b)c)d) |  | a)  If RAN2 can reach consensus on the assumption that the entity of model training and model storage is the same one, rephrase ‘Model training’ to ‘Model training& model storage’  For the mapped entities, the original agreement of RAN1 did not mention OTT server for model training at UE-side. However, due to limited capability, the UE may not be able to perform model training and may offload the training functionality to OTT server. Therefore, the OTT server can be clarified as ‘UE/OTT server’.  Besides, the training type shall be clarified:  For training type 1 (joint training): UE/OTT server (UE side training) or gNB/OAM (NW side training)  For training type 3 (separate training): UE/OTT server and gNB/OAM  b)  Rephrase as:  For training Type 1 UE side training: UE->gNB (if model training in UE) or OTT server->gNB&UE (if model training is offload to OTT server)  For training Type 1 NW side training: gNB->UE, or OAM->gNB&UE  For training Type 3 UE part model: No model transfer/delivery (if model training in UE) or OTT server to UE (if model is trained in OTT server)  For training Type 3 NW part model: No transfer/delivery (if model training in gNB) or OAM to gNB (if model is trained in OAM)  c)  Rephrase as:  UE (CSI generation part) and gNB (CSI reconstruction part)  d)  As only one of NW or UE may monitor the performance, rephrase as:  gNB (NW monitors the performance) or UE (UE monitors the performance and reports to Network)  e)  Agree with the current one as RAN1 only agreed the cases that NW makes decisions of model activation/ deactivation/updating/switching. |
| Xiaomi | All with comments on e |  | In general, we support rapp’s proposals, which captures the current common understanding.  Regarding e) Model/functionality control (selection, (de)activation, switching, fallback), RAN1 made following agreement  ***In CSI compression using two-sided model use case, study potential specification impact for performance monitoring including:***   * ***NW-side performance monitoring: NW monitors the performance and make decisions of model activation/ deactivation/updating/switching***   ***UE-side performance monitoring: UE monitors the performance and reports to Network, NW makes decisions of model activation/ deactivation/updating/switching***  Therefore, we suggest to add ‘updating’ as one potential control action. |
| Mediatek | c) |  | For bullet a) model training, although the OTT server can be a NW-sided server or an UE-sided server, it should be the UE-sided OTT server concerned in the discussion.  Possible revision:  gNB, OAM, UE-sided OTT server  For bullet b), it’s better to differentiate model transfer and model delivery, because model transfer and model delivery have different meaning and specification impact.  For training type 1: Joint training of the two-sided model at a single side/entity, e.g., UE-sided or Network-sided.  Model transfer over air interface may be required either from network side to UE side (gNB->UE)or from UE side to network side (UE->gNB). If the model is trained by the NW at OAM, the model transfer is still from gNB to UE and OAM is invisible to UE.  The description of ‘For training Type 1: gNB->UE, or OAM->gNB&UE, or OTT server->gNB&UE’ is confusing. OAM->gNB&UE/ OTT server->gNB&UE means the two-sided model is delivered from OAM/OTT server to gNB and UE respectively? So gNB->UE means model transfer over air interface and OAM->gNB&UE/ OTT server->gNB&UE means model delivery from OAM/OTT through UP channel? But we don’t have direct model delivery from OAM to UE.  Possible revision:  For training Type 1:  Model transfer gNB->UE (together with model delivery from OAM to gNB if model is trained at OAM); or  Model transfer UE->gNB (together with model delivery from OTT server to UE), or  Model delivery OTT server->gNB&UE respectively.  For training Type 3: For UE-side model, model delivery OTT server->UE if the UE-side model is trained at UE-side OTT server; For NW-side model, no model transfer/delivery if the NW-side model is trained at gNB, or model delivery OAM->gNB if the NW-side model is trained at OAM;  For bullet (d), for model/functionality monitoring, we may have long-term, medium-term and short-term monitoring. For short/medium-term monitoring, it can happen at gNB or UE. For long-term monitoring, it can be performed at OAM or UE-sided OTT server.  For bullet (e), UE autoneouse control and UE report to network with UE monitoring should be allowed at the time being.  For data share/transfer, I agree with Apple that it’s an important aspect to be considered. But it can be considered as one part of data collection and don’t need to be addressed in this email discussion. |
| Lenovo | c) d) are ok  a)b)e) see comments |  | Agree with the terminology comment from Apple, it should be “UE part of two sided model” and “gNB part of two sided model”  a) it would be good to distinguish between UE part model and gNB part model as well. And in our understanding:  - Training of UE part model: gNB/OAM/OTT server  - Training of gNB part mode: gNB/OAM (it’s actually upon RAN3’s decision, maybe RAN2 does not need to discuss this)  b) for training type 1, maybe more precise to say e.g.,  - UE part model: (OAM/OTT 🡪) gNB 🡪 UE  - gNB part model: (OAM 🡪) gNB. No proposal about the OTT 🡪 gNB solution yet. In addition, the option UE provides a trained gNB part model to gNB is also possible, i.e., UE 🡪 gNB.  c) d) ok  e) Again better to distinguish between UE part and NW part model  - UE part: both gNB decided, or UE decided model/functionality control are possible  - gNB part: gNB decided  Some clarification on “Note 1: For a), only data collection part may be further discussed.”. What is the intention? Data collection applies to inference and monitoring as well. |
| CATT | All with comments and modifications |  | For a), we think model can also be trained at UE side, so UE should be added, as “gNB, OAM, OTT server, UE”.  For b), as mentioned in a), for training Type 1, if model is trained at UE side, “UE->gNB” should be added, as “gNB->UE, UE->gNB, or OAM->gNB&UE, or OTT server->gNB&UE”;  for training Type 3, for UE-side model, if model is trained at UE side, “no model transfer/delivery for UE-side model” should be added, and the model training can also at OAM, so propose to “no model transfer/delivery for UE-side model if the UE-side model is trained at UE, OTT server->UE if the UE-side model is trained at OTT server, OAM->UE if the UE-side model is trained at OAM”;  for training Type 3,for NW-side model, we think the model training can also at OTT server, so propose to “For NW-side model, no model transfer/delivery if the NW-side model is trained at gNB, or OAM->gNB if the NW-side model is trained at OAM, or OTT server->gNB if the NW-side model is trained at OTT server;”  In addition, we agree Apple’s suggestion that change “UE side model” to “UE part of two-sided model” and change “NW side mode” to “NW part of two-sided model” to align with RAN1 and the terms in TR 38.843.  For c), we think it is better to be changed to “NW part of two-sided model: gNB, UE part of two-sided model: UE” for accuracy.  For d), it is also better to change to “Model monitoring at NW: gNB, Model monitoring at UE: UE” as “NW-side: gNB, UE-side: UE” is easy to be understanded to gNB monitors NW part of two-sided model, UE monitors UE part of two-sided model, however, the model monitoring is for entire two-sided model not separately.  For e), considering that RAN1 has agreed UE may activate/deactivate/select/switch individual AI/ML models via model ID, “UE” should be added, as “gNB,UE”. |
| Qualcomm | c) | a) b) d) e) | For a, the model training can happen within the CN at NWDAF, AS, AF, and others. There is no reason for excluding CN. Where model training can be performed should be left to vendor choice. Therefore, CN should be considered for model training.  For b, we have option 2, where the model is transferred to the UE/gNB from the core network. Model training and storage can happen at different places. Therefore, model transfer from CN-> gNB/UE should be included for both type 1 and type 3 training.  For d, we would like to wait for RAN1 progress, i.e., whether, how, and where monitoring is performed. Furthermore, we agree with CATT that using NW-side and UE-side may be confusing (i.e., whether NW-side monitoring implies monitoring of network side model). Maybe we can modify by removing side  NW~~-side~~: gNB  UE~~-side~~: UE  For e, based on RAN1#110bis-emeeting agreements below  -------------------  Agreement  For model selection, activation, deactivation, switching, and fallback at least for UE sided models and two-sided models, study the following mechanisms:   * Decision by the network   + Network-initiated   + UE-initiated, requested to the network * Decision by the UE   + Event-triggered as configured by the network, UE’s decision is reported to network   + UE-autonomous, UE’s decision is reported to the network   + UE-autonomous, UE’s decision is not reported to the network   FFS: for network sided models  FFS: other mechanisms  ------------------  Model/functionality control (selection, (de)activation, switching, fallback) can be performed by the UE. |

Summary of Q1:

### 2.1.2 CSI prediction with UE-side model

*Rapporteur’s notes: It is observed that CSI prediction with UE-side model can share similar analysis to beam management with UE-side model. However, considering there is no more discussion in RAN1, it may be better to wait for RAN1’s progress.*

## 2.2 Beam management

RAN1 agreed to support BM-Case1 and BM-Case2 for beam management with one-sided model (i.e. UE-side model or network-side model).

• BM-Case 1: Spatial-domain DL beam prediction

• BM-Case 2: Temporal DL beam prediction

Based on RAN1 progress, the similar mechanism for LCM procedure is used for BM-Case1 and BM-Case2.

### 2.2.1 UE-side model

For beam management with UE-side model, the AI/ML model training and inference both at UE side is at least supported. Model training at NW side and model inference at UE side may be further studied based on the support of model transfer. RAN2 assumed that training data can be generated by UE/gNB and terminated at gNB/OAM/OTT server. It is suggested that model training can reside at gNB/OAM/OTT server, so the model can be transferred/delivered to the UE from gNB/OAM/OTT server.

For model monitoring and control, UE monitors the performance metric and UE make decisions of model control, or UE monitors the performance metric and gNB make decisions of model control are potential solutions based on RAN1 agreements.

|  |
| --- |
| **RAN1#110bis Agreement**  For BM-Case1 and BM-Case2 with a UE-side AI/ML model, study the following alternatives for model monitoring with potential down-selection:   * Atl1. UE-side Model monitoring   + UE monitors the performance metric(s)   + UE makes decision(s) of model selection/activation/ deactivation/switching/fallback operation * Atl2. NW-side Model monitoring   + NW monitors the performance metric(s)   + NW makes decision(s) of model selection/activation/ deactivation/switching/ fallback operation * Alt3. Hybrid model monitoring   + UE monitors the performance metric(s)   + NW makes decision(s) of model selection/activation/ deactivation/switching/ fallback operation   **RAN1#111 Agreement**  For the sub use case BM-Case1 and BM-Case2, at least support Alt.1 and Alt.2 for AI/ML model training and inference for further study:   * Alt.1. AI/ML model training and inference at NW side * Alt.2. AI/ML model training and inference at UE side * The discussion on Alt.3 for BM-Case1 and BM-Case2 is dependent on the conclusion/agreement of Agenda item 9.2.1 of RAN1 and/or RAN2 on whether to support model transfer for UE-side AI/ML model or not * Alt.3. AI/ML model training at NW side, AI/ML model inference at UE side |

The possible mapping of AI/ML functions to physical entities for beam management with a UE-side model is shown in the following table.

Table 2.2-1: The mapping of AI/ML functions to physical entities for beam management with UE-side model

|  |  |  |
| --- | --- | --- |
|  | **AL/ML functions (if applicable)** | **Mapped entities** |
| a) | Model training | gNB, OAM, OTT server |
| b) | Model transfer/delivery | gNB->UE, or OAM->UE, or OTT server->UE |
| c) | Inference | UE |
| d) | Model/functionality monitoring | UE, gNB |
| e) | Model/functionality control (selection, (de)activation, switching, fallback) | gNB if monitoring resides at UE or gNB,  UE if monitoring resides at UE |

Note 1: For a), only data collection part may be further discussed.

Note 2: Whether/how OAM is to be invovled may need to consult SA5.

Q2: Do you agree the mapping of functions to physical entities for beam management with UE-side model in Table 2.2-1?

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Yes for which bullet(s)** | **No for which bullet(s)** | **Comments** |
| Apple | a), b), c), d), e) | dataset transfer is missed | Similar to Q1, we think dataset transfer is missed. Thus, we suggest to add row f)  f): dataset transfer:   * For training: from UE to OTT server (if model is trained in OTT server), or from UE to gNB (if model is trained in gNB), or from UE to OAM (if model trained in OAM) * For Monitoring: if NW monitors, from UE to gNB. |
| OPPO | All with comments and modification |  | For bullet a, based on RAN1 agreements above, UE side model can be trained at UE, so this scenario should be considered also.  For bullet b, the missing part is that no model transfer/delivery if the UE-side model is trained at UE.  For bullet d, it’s not very clear how to differentiate the three Alternatives given by RAN1, so we suggest:  UE if UE monitors the performance metrics, UE or gNB makes the decision of model selection/activation/ deactivation/switching/fallback operation.  gNB if gNB monitors the performance metrics. |
| Mavenir | a), b), c), d), e) |  | we agree with OPPO’ s opinion |
| vivo | All with comments on a)b)d) |  | a)  See above in Q1, the OTT server can be clarified as ‘UE/OTT server’.  b)  Rephrase as:  For training at NW side: gNB->UE, or OAM-> UE  For training at UE side: No model transfer/delivery (if model training in UE) or OTT server to UE (if model is trained in OTT server)  d)  Rephrase as:  gNB (NW monitors the performance) or UE (UE monitors the performance and reports to Network) |
| Xiaomi | All with commet on b |  | Since Model training at NW side and model inference at UE side is not supported yet in RAN1, we propose to make corresponding model delivery as FFS to align with RAN1. Following change is suggesuted,  [gNB->UE, or OAM->UE], OTT server->UE’. |
| Mediatek | c), e) |  | For bullet a) model training, same as Q1, the discussion concerns UE-sided OTT server.  For bullet b), still don’t know how model transfer/delivery from OAM to UE happen. We need to differentiate model delivery and model transfer  Possible revision:  [Model transfer: gNB->UE], or  Model delivery: [OAM->UE, or] UE-side OTT server->UE  For bullet d), same comment as Q1, For short/medium-term monitoring, it can happen at gNB or UE. For long-term monitoring, it can be performed at OAM or UE-sided OTT server. |
| Lenovo | All, with comments |  | Model training by UE itself should also be considered.  Clarification as commented by other companies could helpful.  Also, would be good to clarify the intention of “Note 1: For a), only data collection part may be further discussed.” |
| CATT | b), c) , d), e)  and a) with modification |  | For a) we share the same view that model training can be performed at UE side. So it can be revised as:  UE, gNB, OAM, OTT server; |
| Qualcomm | c) | a)b)d)e) | For a, the model training can happen within the CN at NWDAF, AS, AF, and others. There is no reason for excluding CN. Where model training can be performed should be left to vendor choice. Therefore, CN should be considered for model training.  For b, we have options 2a and 2b, where the model is transferred to the UE/gNB from the core network. Model training and storage can happen at different places. Therefore, mode transfer from CN-> UE should be included.  For d, we would like to wait for RAN1 progress, i.e., whether, how, and where monitoring is performed.  For e, based on RAN1#110bis-emeeting agreements below  -------------------  Agreement  For model selection, activation, deactivation, switching, and fallback at least for UE sided models and two-sided models, study the following mechanisms:   * Decision by the network   + Network-initiated   + UE-initiated, requested to the network * Decision by the UE   + Event-triggered as configured by the network, UE’s decision is reported to network   + UE-autonomous, UE’s decision is reported to the network   + UE-autonomous, UE’s decision is not reported to the network   FFS: for network sided models  FFS: other mechanisms  ------------------  We have a RAN1 agreement for this. The following are not required:  gNB ~~if monitoring resides at UE or gNB~~,  UE ~~if monitoring resides at UE~~ |

Summary of Q2:

### 2.2.2 NW-side model

For beam management with NW-side model, model inference is naturally at gNB side. For model training, the model can be trained at gNB or OAM side, similarly to the mechanism studied in RAN3 AI for NG-RAN. If the model is trained at OAM side, model transfer/delivery is needed from OAM to gNB.

For model/functionality monitoring and control, gNB monitors the performance metric(s) and makes decision(s) of control is supported based on RAN1 agreements.

|  |
| --- |
| **RAN1#110bis Agreement**  For BM-Case1 and BM-Case2 with a network-side AI/ML model, study the NW-side model monitoring:   * NW monitors the performance metric(s) and makes decision(s) of model selection/activation/ deactivation/switching/ fallback operation |

The possible mapping of AI/ML functions to physical entities for beam management with a NW-side model is shown in the following table. For NW-side model, it seems that only data collection (e.g. for model training, inference, monitoring, control) has RAN2 impacts, and other LCM purposes can be up to NW implementation.

Table 2.2-2: The mapping of functions to physical entities for beam management with NW-side model

|  |  |  |
| --- | --- | --- |
|  | **AL/ML functions (if applicable)** | **Mapped entities** |
| a) | Model training | gNB, OAM |
| b) | Model transfer/delivery | OAM->gNB, or N/A if the model is trained at gNB |
| c) | Inference | gNB |
| d) | Model/functionality monitoring | gNB |
| e) | Model/functionality control (selection, (de)activation, switching, fallback) | gNB |

Note 1: For NW-sided model, only data collection part may be further discussed.

Note 2: Whether/how OAM is to be involved may need to consult SA5.

Q3: Do you agree the mapping of functions to physical entities for beam management with NW-side model in Table 2.2-2?

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Yes for which bullet(s)** | **No for which bullet(s)** | **Comments** |
| Apple | a), b), c), d), e) | dataset transfer is missed | Similar to Q1/2, we think dataset transfer is missed. Thus, we suggest to add row f)  f): dataset transfer:   * For training: from gNB to OAM (if model trained in OAM) |
| OPPO | All with comments and modification |  | For bullet a, based on RAN2 agreements, OTT server is also one candidate:  For CSI enhancement and beam management use cases:  - For model training, training data can be generated by UE/gNB and terminated at gNB/OAM/OTT server.  So we just wonder why OTT server case is not included, my understanding is that this OTT server can be an operator trusted node to train network side model.  For bullet b, based on above comments for bullet a, we have the following suggestion:  OAM->gNB, OTT server to gNB, or no model transfer/delivery if the model is trained at gNB |
| Mavenir | a), b), c), d) , e) |  | We are ok with OPPO’ s suggestion about a) b) |
| vivo | All with comments on b) |  | b)  further clarified as:  No model transfer/delivery (if model training in gNB) or OAM to gNB (if model training in OAM) |
| Xiaomi | All |  |  |
| Mediatek | a),c),d),e) |  | Possible revision:  Model delivery: OAM->gNB, or  N/A if the model is trained at gNB |
| Lenovo | All, with comment |  | Ok to us, although it may not be upon RAN2 decision at the end where a gNB side model will be trained. |
| CATT | c), d), e)  and a), b) with comments |  | 1. We share the save view that OTT server should be included. So it can be revised as:   gNB, OAM, OTT server.   1. It can be revised as:   OAM->gNB, or N/A if the model is trained at gNB, OTT server->gNB |
| Qualcomm | c) | a) b) d) e) | For a, the CN and OTT server should be included.  For b, the model delivery from OTT server-> gNB, and CN-> gNB should be included.  For d, monitoring (at least long term) can be performed at the OAM.  For e) OAM can perform LCM for model running at gNB. |

Summary of Q3:

## 2.3 Positioning accuracy enhancement

RAN1 agreed the following cases for AI/ML-based positioning accuracy enhancement, which can be categorized to three model types, i.e. UE-side model, LMF-side model and gNB-side model.

* Case 1: UE-based positioning with UE-side model, direct AI/ML or AI/ML assisted positioning
* Case 2a: UE-assisted/LMF-based positioning with UE-side model, AI/ML assisted positioning
* Case 2b: UE-assisted/LMF-based positioning with LMF-side model, direct AI/ML positioning
* Case 3a: NG-RAN node assisted positioning with gNB-side model, AI/ML assisted positioning
* Case 3b: NG-RAN node assisted positioning with LMF-side model, direct AI/ML positioning

### 2.3.1 UE-side model

For case 1 and 2a with UE-side model, RAN2 assumed that training data can be generated by UE/gNB and terminated at LMF/OTT server. Thus, model training at LMF/OTT server is feasible. For model monitoring, at least UE and LMF can derive monitoring metric per RAN1 agreement. The decision of model control can be also made at least by UE or LMF.

|  |
| --- |
| **RAN1#112bis-e Agreement**  Regarding monitoring for AI/ML based positioning, at least the following entities are identified to derive monitoring metric   * UE at least for Case 1 and 2a (with UE-side model) * gNB at least for Case 3a (with gNB-side model) * LMF at least for Case 2b and 3b (with LMF-side model)   **RAN1#113 Agreement**  Regarding AI/ML model monitoring for AI/ML based positioning, the following entities are identified as candidates to derive monitoring metric in addition to entities from previous agreement   * LMF for Case 2a (with UE-side model) and Case 3a (with gNB-side model) at least when monitoring is based on provided ground truth label (or its approximation) |

The mapping of AI/ML functions to physical entities for case 1 and 2a with UE-side model is list in the following table.

Table 2.3-1: The mapping of functions to physical entities for positioning with UE-side model (case 1 and 2a)

|  |  |  |
| --- | --- | --- |
| **Use case** | **AL/ML functions (if applicable)** | **Mapped entities** |
| a) | Model training | LMF, OTT server |
| b) | Model transfer/delivery | LMF->UE, or OTT server->UE |
| c) | Inference | UE |
| d) | Model/functionality monitoring | UE, LMF |
| e) | Model/functionality control (selection, (de)activation, switching, fallback) | UE if monitoring resides at UE,  LMF if monitoring resides at UE or LMF |

Note 1: For a), only data collection part may be further discussed.

Note 2: Whether/how OAM is to be involved may need to consult SA5.

Note 3: Whether/how LMF is to be involved may need to consult SA2.

Q4: Do you agree the mapping of functions to physical entities for positioning with UE-side model (case 1 and 2a) in Table 2.3-1?

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Yes for which bullet(s)** | **No for which bullet(s)** | **Comments** |
| Apple | a), b), c), d), e) | dataset transfer is missed | Similar to Q1, we think dataset transfer is missed. Thus, we suggest to add row f)  f): dataset Transfer:   * For training: from UE/PRU to OTT server (if model is trained in OTT server), or from UE/PRU to LMF (if model is trained in LMF). * For Monitoring: if LMF monitors, from UE to LMF.   Another issue is whether OAM can be used for training purpose (related to Note 2). We think it can be one alternative for UE-side model but open for discussion. In that case, we can update:  a) LMF, OTT server, OAM  b) LMF->UE, or OTT server->UE, or OAM->UE |
| OPPO | All |  |  |
| Mavenir | a), b), c), d), e) |  | We are ok with Apple’ s suggestion about a) b) |
| vivo | All with comments on a)b)d) |  | a)  See above in Q1, the OTT server can be clarified as ‘UE/OTT server’.  b)  Rephrase as:  For training at NW side: LMF->UE  For training at UE side: No model transfer/delivery (if model training in UE) or OTT server to UE (if model is trained in OTT server)  d)  Rephrase as:  LMF (NW monitors the performance) or UE (UE monitors the performance) |
| Xiaomi | All with comment on a and b |  | Seems OAM is excluded. No strong opinion, but we wonder why to exclude OAM training model. |
| Mediatek | c), d), e) |  | For bullet a) model training, same as Q1, the discussion concerns UE-sided OTT server.  For bullet b), We need to differentiate model delivery and model transfer  Possible revision:  [Model transfer: LMF->UE], or  Model delivery: UE-side OTT server->UE |
| Lenovo | All, with comment |  | Model training by UE itself should also be considered.  Also, would be good to clarify the intention of “Note 1: For a), only data collection part may be further discussed.” |
| CATT | All |  | 1. We think the training can also be performed by the UE, OAM. (UE is not excluded for model training based on RAN1 agreement, and whether OAM can be used for training for UE-side model can be discussed.)   May add OAM->UE. |
| Qualcomm | c) | a)b)d)e) | For a, the model training can happen within the CN at NWDAF, AS, AF, and others. There is no reason for excluding CN. Where model training can be performed should be left to vendor choice. Therefore, CN should be considered for model training.  For b, we have options 2a and 2b, where the model is transferred to the UE from the core network. Model training and storage can happen at different places. Therefore, mode transfer from CN-> UE should be included.  For d, we would like to wait for RAN1 progress, i.e., whether, how, and where monitoring is performed.  For e, based on RAN1#110bis-emeeting agreements below  -------------------  Agreement  For model selection, activation, deactivation, switching, and fallback at least for UE sided models and two-sided models, study the following mechanisms:   * Decision by the network   + Network-initiated   + UE-initiated, requested to the network * Decision by the UE   + Event-triggered as configured by the network, UE’s decision is reported to network   + UE-autonomous, UE’s decision is reported to the network   + UE-autonomous, UE’s decision is not reported to the network   FFS: for network sided models  FFS: other mechanisms  ------------------  We have a RAN1 agreement for this. The following are not required:  UE ~~if monitoring resides at UE~~,  LMF ~~if monitoring resides at UE or LMF~~ |

Summary of Q4:

### 2.3.2 LMF-side model

For case 2b and 3b with LMF-side model, it is straightforward that model training and inference are both at LMF side. In this case, model transfer/delivery is not needed. For monitoring and control, RAN1 agreed that at least LMF can derive monitoring metric and make decisions of control.

The mapping of AI/ML functions to physical entities for case 2b and 3b with LMF-side model is list in the following table. For LMF-side model, it seems that only data collection (e.g. for model training, inference, monitoring, control) has spec impacts, and other LCM purposes can be up to NW implementation.

Table 2.3-2: The mapping of functions to entities for positioning with LMF-side model (case 2b and 3b)

|  |  |  |
| --- | --- | --- |
|  | **AL/ML functions (if applicable)** | **Mapped entities** |
| a) | Model training | LMF |
| b) | Model transfer/delivery | N/A |
| c) | Inference | LMF |
| d) | Model/functionality monitoring | LMF |
| e) | Model/functionality control (selection, (de)activation, switching, fallback) | LMF |

Note 1: For LMF-side model, only data collection part may be further discussed.

Note 2: Whether/how OAM is to be involved may need to consult SA5.

Note 3: Whether/how LMF is to be involved may need to consult SA2.

Q5: Do you agree the mapping of functions to physical entities for positioning with LMF-sided model (case 2b and 3b) in Table 2.3-2?

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Yes for which bullet(s)** | **No for which bullet(s)** | **Comments** |
| Apple | a), b), c), d), e) | dataset transfer is missed | Similar to Q1, we think dataset transfer is missed. Thus, we suggest to add row f)  f): dataset transfer:   * For training: from UE/PRU to LMF (case 2b), from gNB to LMF (case 3b). * For Inferecne: from UE/PRU to LMF (case 2b), from gNB to LMF (case 3b). * For Monitoring: from UE/PRU to LMF (case 2b), from gNB to LMF (case 3b).   Another issue is whether OAM can be used for training purpose (related to Note 2). We think it may not a good solution for LMF-sided model because there seems no interface between LMF and OAM. If majority also think it can't, Note 2 can be removed. |
| OPPO | All |  |  |
| Mavenir | All |  |  |
| vivo | All |  |  |
| Xiaomi | All with comment on a and b |  | Seems OAM is excluded. No strong opinion, but we wonder why to exclude OAM training model. |
| Mediatek | all |  |  |
| Lenovo | All |  | OK to us, although it may not be upon RAN2 decision at the end where a LMF side model is trained. |
| CATT | All |  | 1. We think the training can also be performed by the OTT server, OAM; 2. May add OTT server ->LMF, OAM->LMF,   For d) and e), the monitoring entity or the Model/functionality control entity can also be: UE (for case 2b) or gNB (for case 3b). |
| Qualcomm | c)d)e) | a)b) | For a, OTT server and CN should be included as model training entities.  For b, add other CN entities ->LMF |

Summary of Q5:

### 2.3.3 gNB-side model

For case 3a with gNB-side model, model training can reside at gNB, OAM or LMF side. If model is not trained at gNB side, model transfer/delivery is needed. For monitoring, gNB and LMF can derive monitoring metric based on RAN1 agreements. For model/functionality control, it is feasible that gNB or LMF makes decisions based on monitoring metric.

The mapping of AI/ML functions to physical entities for case 3a with gNB-side model is list in the following table. For gNB-side model, it seems that only data collection (e.g. for model training, inference, monitoring, control) has spec impacts, and other LCM purposes can be up to NW implementation.

Table 2.3-3: The mapping of AI/ML functions to entities for positioning with gNB-side model (case 3a)

|  |  |  |
| --- | --- | --- |
| **Use case** | **AL/ML functions (if applicable)** | **Mapped entities** |
| a) | Model training | gNB, OAM, LMF |
| b) | Model transfer/delivery | LMF-> gNB, or OAM->gNB, or N/A if the model is trained at gNB |
| c) | Inference | gNB |
| d) | Model/functionality monitoring | gNB, LMF |
| e) | Model/functionality control (selection, (de)activation, switching, fallback) | gNB, LMF |

Note 1: For gNB-side model, only data collection part may be further discussed.

Note 2: Whether/how OAM is to be involved may need to consult SA5.

Note 3: Whether/how LMF is to be involved may need to consult SA2.

Q6: Do you agree the mapping of functions to physical entities for positioning with gNB-side model (case 3a) in Table 2.3-3?

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Yes for which bullet(s)** | **No for which bullet(s)** | **Comments** |
| Apple | a), b), c), d), e) | dataset transfer is missed | Similar to Q1, we think dataset transfer is missed. Thus, we suggest to add row f)  f): dataset transfer:   * For training: from gNB to OAM (if model is trained in OAM), from gNB to LMF (if model is trained in LMF) * For Monitoring: from gNB to LMF (if LMF performs monitoring). |
| OPPO | All |  |  |
| Mavenir | All |  |  |
| vivo | All |  |  |
| Xiaomi | All |  |  |
| Mediatek | All |  |  |
| Lenovo | All |  | OK to us, although it may not be upon RAN2 decision at the end where a gNB side model is trained. |
| CATT | All |  | 1. We think the training can also be performed by the OTT server;   May add OTT server->gNB. |
| Qualcomm | c) d) e) | a) b) | For a, OTT server and CN should be included as model training entities.  For b), add CN->gNB |

Summary of Q6:

# 3 Conclusion

To be added...

# 4 Reference

1. R2-2305613 Discussion on general architecture for AI/ML for NR air interface CMCC