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

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

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

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) |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

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 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

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

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). |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

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