3GPP TSG-RAN WG2 #124 R2-23xxxxx

Chicago, USA, 13 – 17 Nov, 2023

Agenda Item: 7.16.2.2

Source: Intel Corporation

Title: [Post123bis][016][AIML] Model Transfer (Intel)

Document for: Discussion, Decision

# Introduction

This document is to address the following email discussion:

* [POST123bis][016][AI/ML] Model transfer (Intel)

Scope: Discuss table that captures pros, cons and specification efforts for the 4 solutions.

Intended outcome: Agreeable proposal/table

Deadline: Nov. 1st

To facilitate the discussion, below agreements from previous meetings related to model transfer are captured:

* RAN2 #120

|  |
| --- |
| * For model transfer/delivery for AI/ML models (for the target use cases of this SI), RAN2 to study CP-based, UP-based solutions |

* RAN2 #121

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| * We Use the wording “model transfer/delivery” * model delivery that serves the use cases in the SI is within RAN2 scope, regardless other aspects. * Agreed:   Aim to at least analyze the feasibility and benefits of model/transfer solutions based on the following:  Solution 1a: gNB can transfer/deliver AI/ML model(s) to UE via RRC signalling.  Solution 2a: CN (except LMF) can transfer/deliver AI/ML model(s) to UE via NAS signalling.  Solution 3a: LMF can transfer/deliver AI/ML model(s) to UE via LPP signalling.  Solution 1b: gNB can transfer/deliver AI/ML model(s) to UE via UP data.  Solution 2b: CN (except LMF) can transfer/deliver AI/ML model(s) to UE via UP data.  Solution 3b: LMF can transfer/deliver AI/ML model(s) to UE via UP data.  Solution 4: Server (e.g. OAM, OTT) can transfer/delivery AI/ML model(s) to UE (e.g. transparent to 3GPP).  **Table: relations between solutions and use cases**   |  |  | | --- | --- | | **Solutions** | **Applicable use cases** | | Solution 1a, 1b | CSI feedback enhancement  Beam management  Note: No specific considerations for Positioning accuracy enhancement for Solution 1a and 1b. | | Solution 2a, 2b | CSI feedback enhancement  Beam management  Note: No specific considerations for Positioning accuracy enhancement for Solution 2a and 2b. | | Solution 3a, 3b | Positioning accuracy enhancement | | Solution 4 | CSI feedback enhancement  Beam management  Positioning accuracy enhancement |   Note: the solutions use case relation is preliminary (work in progress), and the purpose is to have better understanding on what to further analyse   |  |  |  | | --- | --- | --- | |  | **Pros** | **Cons** | | **Solution 1a** | 6. The existing RRC signaling solutions can be reused as baseline, at least including delta signaling and segementation  9. Additional security and verification may not be necessary as the UE already established security before the transfer is initiated  11. gNB can take the control of the AIML model transfer itself, which can not be achieved by traditional UP based solution | 1. Face challenges to convey large size or “no upper limit size” AI model by RRC message (e.g. >45kBytes)  2. Maybe high control plane overhead, as a large model size may need segmentation/transmission/acknowledgment. This consumes critical configuration time for model transfer/delivery  3. An incomplete control plane model transfer has to be restarted upon mobility, as there are no current procedures to resume transmission across gNBs. Some companies wonder whether it is critical or not as it depends on how frequent the gNB to send new/updated AI/ML to the UE | | **Solution 2a and 3a** | 5. Service continuity on model transfer/delivery is easy to achieve compared with Solution 1a  6. Impacts on RAN2 may be limited (some companies think that LPP signalling is in RAN2 scope) | 1. Face challenges to convey large size or “no upper limit size” AI model by RRC message (e.g. >45kBytes)  3. If NAS does the segmentation, it may introduce some overhead  4. (only valid for Solution 2a) 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 | | **Solution 1b** | 1. The network can provide different 5QIs for model transfer/delivery with different QoS requirements (e.g. can support large model size)  2. Compared with CP-based solutions, this Solution 1b can reduces control plane overhead, reduces overhead at gNB for model delivery/transfer  5. Compared with CP-based solutions, it may not need to consider CP message segmentation, CP message blocking issue | 5. Not compatible with current mobility procedure. Supporting model transfer during mobility is not so straightforward | | **Solution 2b and 3b** | 1. The network can provide different 5QIs for model transfer/delivery with different QoS requirements (e.g. can support large model size)  5. Compared with CP-based solutions, it may not need to consider CP message segmentation, CP message blocking issue | 2. CP signalling is needed to configure and initiate the model transfer from the CN  4. May be unable to support delta-model transfer/delivery based on current user plane framework | | **Solution 4** | 2. If 3GPP network can be aware of AI/ML model in this Solution 4, the network can provide different 5QIs for model transfer/delivery with different QoS requirements (e.g. can support large model size). How to synchronize 3GPP and server so that the network can take appropriate actions is not clear, and it may not be fully under 3GPP control | 2. There may be inter-operability issues, such as:  a) Different implementations may lead to different model performances and a huge burden of model management (e.g., frequent model activation/deactivation)  b) Massive offline coordination is needed or requires lots of coordinations among vendors, especially for the CSI compression use case  4. When network cannot control the model transfer/delivery, the transfer of large model may impact important and delay sensitive user data traffic |  * The table can serve as starting point for continued discussion (but contains some parts that seems non consensus, e.g. delta configuration). |

* RAN2 #123

|  |
| --- |
| * Model transfer/delivery can be initiated in following two ways:   Reactive model transfer/delivery: an AI/ML model is downloaded when it is needed due to changes in scenarios, configurations, or sites.  FFS: Proactive model transfer/delivery: AI/ML models are pre-download to UE, and a model switch is performed when changes in scenarios, configurations, or sites occur. |

* RAN2 #123bis

|  |
| --- |
| **=> Agree to split**  ***- Solution 4a: OTT server can transfer/delivery AI/ML model(s) to UE (transparent to 3GPP).***  ***- Solution 4b: OAM can transfer/delivery AI/ML model(s) to UE.*** |

As observed from previous discussion, model transfer is also related to the mapping of functions to entities that were discussed in R2-2308286 [2] and agreed in RAN2 #123:

|  |
| --- |
| [R2-2308286](http://www.3gpp.org/ftp//tsg_ran/WG2_RL2/TSGR2_123/Docs//R2-2308286.zip) Report of [Post122][060][AIML] Mapping of functions to physical entities (CMCC) CMCC report Rel-18 FS\_NR\_AIML\_air  - Quite long discussion  - CMCC report that FFS items has support from 3 companies.  - Chair Comment: These options represent several possibilities. RAN2 would typically have selected a specific architecture option, and for a WI, specific option(s) need to be selected. Hope it is possible to further narrow down during the SI.   * P1-P6 are agreed, it is expected that FFS items for which support is not increased will be removed. |

# Discussion

As mentioned above, this discussion is mainly focuses on pros/cons/spec impact of the four model transfer solutions. Rapporteur understands that there’s a relationship between model transfer and functionality mapping to entities. However, **there’s no intention from this email discussion to down-select among the four model transfer solutions**. Therefore, in below discussion, it is assumed that FFS (e.g. CN, OAM) in mapping of functions to physical entities are considered.

## 2.1 Model Transfer/Delivery Discussion Area

It is observed from the table summarized in R2-2302268 [1] that there are several common areas discussed when comparing different model transfer/delivery solutions, e.g. model size, etc.

Rapporteur believes that summarizing discussion areas for model transfer/delivery could help companies to share the same understanding when discussing feasibility and gap of certain model transfer/delivery solution. Following discussion areas are currently mentioned in the existing table summarized in R2-2302268 [1]:

**A1.** **Large, no upper limit model size** (mentioned in Solution 1a, Solution 2a and 3a, Solution 1b, Solution 4)

It is observed that all solutions can support model transfer/delivery with model size smaller than 45kBytes by default, e.g. CP solutions can support model size smaller than 45kBytes based on existing number of RRC segments. Therefore, for this discussion area, we only focus on model transfer/delivery for model size larger than 45kBytes.

**A2. Security and integrity** (mentioned in Solution 1a)

**A3. Latency requirement, e.g. critical, relax, no latency requirement** (mentioned in Solution 2a)

It is observed that air interface latency can be the same for all solutions by proper setting (e.g. priority setting for SRB/DRB, etc). Therefore, we only focus on the delta latency component for each solution.

**A4. Model transfer/delivery continuity (i.e. resume transmission of model (segments) across gNBs)** (mentioned in Solution 1a, Solution 2a, Solution 1b)

**A5. NW controllability (e.g. model management decision at gNB)** (mentioned in Solution 1a, Solution 2a)

**A6. Partial model update (e.g. delta configuration)** (mentioned in Solution 1a, Solution 2b and 3b)

**A7. Flexible model transfer/delivery QoS** (mentioned in Solution 1b, Solution 2b and 3b, Solution 4)

Different models allow to use different QoS

**A8. Interoperability (e.g. No/minor need for offline coordination among vendors)** (mentioned in Solution 4)

**Q1: Do you agree the above discussion areas should be considered during discussion of model transfer/delivery solutions?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No (please list the item(s) correspondingly)** | **Comment** |
| OPPO | No at least for A2/A8 | A2  No matter CP or UP based solution is considered, Security and integrity is already supported in legacy, so no need to consider this as the pros or cons for any specific solution.  A8  If open model format is used for model transfer/delivery solution1a, there is no inter-operability issue as all devices can recognize the details of the open format model. If proprietary model format is used for model transfer/delivery solution1a, inter-operability issue may happen as usually one vendor cannot recognize the details of the proprietary format model from another vendor. But it should be noted that this restriction is not only applied to model transfer/delivery solution1a, but also applied to all the other model transfer/delivery solutions. In this sense, we can know that solution1a has no advantage over the other solutions on inter-operability aspect.  **Observation:Model transfer/delivery solution1a has no advantage over the other model transfer/delivery solutions on inter-operability aspect.**  It does not make sense to consider inter-operability aspect as one of the pros or cons for a specific model transfer/delivery solution when evaluating each candidate solution, so we propose the following:  **Proposal: Do not consider inter-operability aspect when evaluating pros and cons for each candidate solution.** |
|  |  |  |
|  |  |  |
|  |  |  |

## 2.2 Model Transfer/Delivery Table

According to Rapporteur’s observation, during past discussion, companies may have different understanding on whether one bullet is a pro or con for certain solution, considering pros/cons are quite strong wording. Therefore, it is proposed to update the table by stating the objective facts, rather than stating pros/cons in a subjective position.

With the listed discussion areas above, “pros” used in existing table can be considered as “**readiness** of solution x”, indicating **which discussion area(s) are considered (or not) by certain solution**. On the other hand, “cons” used in existing table can be considered as “**current status and gaps**”, by **stating current supporting level and gaps for “not support” discussion area**. In the end, “RAN specification impact” includes potential RAN specification impact. It should be clear that “**RAN specification impact**” column tends to **identify area of specification impact**, rather than being a complete collection of detailed solution.

Furthermore, regarding to CN impact, Rapporteur understands that some solutions may have CN impact. However, it is not RAN2 scope to discuss the details of what CN impact is. Therefore, Rapporteur proposes that, for those solutions that may have potential CN impact, instead of going into details during this email discussion, we add a Note in “RAN specification impact” column by stating there may be a CN impact and leaving the details to SA during normative phase.

With above considerations, please find below mapping between original context to discussion area as below:

|  |  |  |
| --- | --- | --- |
|  | **Pros** | **Cons** |
| **Solution 1a** | 6. The existing RRC signaling solutions can be reused as baseline, at least including delta signaling and segementation  => A6: support partial model update based on RRC delta signaling  9. Additional security and verification may not be necessary as the UE already established security before the transfer is initiated  => A2: security and verification is supported, as security is established by UE based on existing procedure  11. gNB can take the control of the AIML model transfer itself, which can not be achieved by traditional UP based solution  => A5: gNB can control management directly, no additional interaction between management and model transfer is needed over NW interfaces | 1. Face challenges to convey large size or “no upper limit size” AI model by RRC message (e.g. >45kBytes)  2. Maybe high control plane overhead, as a large model size may need segmentation/transmission/acknowledgment. This consumes critical configuration time for model transfer/delivery  => A1: model size >45kBytes is not supported based on existing number of RRC segments  3. An incomplete control plane model transfer has to be restarted upon mobility, as there are no current procedures to resume transmission across gNBs.  => A4: transmission is restarted upon mobility  Some companies wonder whether it is critical or not as it depends on how frequent the gNB to send new/updated AI/ML to the UE  => A3: procedure latency depends on model size and SRB priority |
| **Solution 2a and 3a** | 5. Service continuity on model transfer/delivery is easy to achieve compared with Solution 1a  => A4: For Solution 2a, support within AMF coverage area based on PDCP status report; For Solution 3a, support within LMF coverage area based on LPP signaling segmentation  6. Impacts on RAN2 may be limited (some companies think that LPP signalling is in RAN2 scope)  => Note: The details of model transfer/delivery procedure from CN to UE is out of RAN scope. | 1. Face challenges to convey large size or “no upper limit size” AI model by RRC message (e.g. >45kBytes)  3. If NAS does the segmentation, it may introduce some overhead  => A1: model size >45kBytes is not supported based on existing number of RRC segments  4. (only valid for Solution 2a) 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  => A5: For Solution 2a, gNB cannot perform management directly, considering model transfer is transparent to gNB. |
| **Solution 1b** | 1. The network can provide different 5QIs for model transfer/delivery with different QoS requirements (e.g. can support large model size)  => A7: support by existing QoS-DRB mapping  2. Compared with CP-based solutions, this Solution 1b can reduces control plane overhead, reduces overhead at gNB for model delivery/transfer  => A1: support different model sizes  5. Compared with CP-based solutions, it may not need to consider CP message segmentation, CP message blocking issue  => A1: support different model sizes | 5. Not compatible with current mobility procedure. Supporting model transfer during mobility is not so straightforward  => A4: No solution support model transfer/delivery service continuity if DRB terminated at gNB |
| **Solution 2b and 3b** | 1. The network can provide different 5QIs for model transfer/delivery with different QoS requirements (e.g. can support large model size)  => A7: support by existing QoS-DRB mapping  5. Compared with CP-based solutions, it may not need to consider CP message segmentation, CP message blocking issue  => A1: support different model sizes | 2. CP signalling is needed to configure and initiate the model transfer from the CN  => A5: gNB cannot perform model management directly, NAS signalling is used to configure and initiate model transfer from CN.  4. May be unable to support delta-model transfer/delivery based on current user plane framework  => A6: CN cannot support delta-model transfer/delivery in user plane |
| **Solution 4** | 2. If 3GPP network can be aware of AI/ML model in this Solution 4, the network can provide different 5QIs for model transfer/delivery with different QoS requirements (e.g. can support large model size).  => A7: support different 5QIs  => A1: support different model sizes  How to synchronize 3GPP and server so that the network can take appropriate actions is not clear, and it may not be fully under 3GPP control | 2. There may be inter-operability issues, such as:  a) Different implementations may lead to different model performances and a huge burden of model management (e.g., frequent model activation/deactivation)  b) Massive offline coordination is needed or requires lots of coordinations among vendors, especially for the CSI compression use case  => A8 is not supported  4. When network cannot control the model transfer/delivery, the transfer of large model may impact important and delay sensitive user data traffic  => A5 is not supported |

Besides, Rapporteur also provides additional information based on understanding from previous discussion on readiness and current status/gaps, which are marked in grey.

With that, the updated tables can be found as below. To facilitate this email discussion, rapporteur splits the table under different subsections for each solution.

##### Solution 1a: gNB can transfer/deliver AI/ML model(s) to UE via RRC signalling

Table . Solution 1a Readiness and RAN specification impact

|  |  |  |
| --- | --- | --- |
| **Discussion Area** | **Readiness** | **RAN specification impact** |
| **Current status and Gaps** |
| A1 | not supported ☹ | extension of the number of RRC segments is required to support models larger than 45kBytes |
| maximum 45kBytes based on existing number of RRC segments |
| A2 | supported 😊 |  |
| A3 | procedure latency depends on model size and SRB priority |  |
| A4 | not supported☹ | Introduce service continuity support for SRBs |
| transmission is restarted upon mobility |
| A5 | supported 😊 |  |
| A6 | supported 😊 |  |
| A7 | not supported ☹ | introduce multiple SRBs or SRB with variable/multiple priorities |
| SRB priority is used |
| A8 | supported 😊 |  |

**Q2-1a: For Solution 1a, do you agree the content in above table to capture analysis of model transfer/delivery solutions? (please include comments only on the current context in this question. For new add-ons, please see Q3-1a)**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comment** |
| #example | Yes: Ax No: Ay | Ax:  Ay: |
| OPPO | No  A2/A6/A8 | A2: See comments in Q1;  A6:  Working Assumption in RAN1#111 meeting  *Consider “proprietary model” and “open-format model” as two separate model format categories for RAN1 discussion,*   |  |  | | --- | --- | | *Proprietary-format models* | *ML models of vendor-/device-specific proprietary format, from 3GPP perspective*  *NOTE: An example is a device-specific binary executable format* | | *Open-format models* | *ML models of specified format that are mutually recognizable across vendors and allow interoperability, from 3GPP perspective* |   *From RAN1 discussion viewpoint, RAN1 may assume that:*   * *Proprietary-format models are not mutually recognizable across vendors, hide model design information from other vendors when shared.* * *Open-format models are mutually recognizable between vendors, do not hide model design information from other vendors when shared*   **Based on above RAN1 agreement and working assumption, we can know both open format and proprietary-format are considered for model transfer:**  **In legacy, delta signaling is only applied to control plane in DL.**  When it comes to model transfer/delivery, we think the meaning is totally different. It’s not about control plane parameter update, but about delta model update. Usually AI/ML model algorithm data can be divided into two parts, i.e. model algorithm structure parameters and model algorithm weight parameters. If only model algorithm weight parameters are changed and the model algorithm structure parameters are known by the model receiver, delta model update can be considered to only update model algorithm weight parameters without changing model algorithm structure parameters. This is the typical scenario for delta model update, which can save the signaling overhead for model update procedure especially when the whole model size is very big. But it seems that this scenario is only applied to open format model case as the gNB can recognize the details of the AI/ML model algorithm, so it’s possible to define two separate parameters for an open format model, one is for model algorithm structure parameters while another is for model algorithm weight parameters. In this way, delta model update can be achieved like the legacy way used for delta signaling. But this delta model update definition is only applied to open format model case, if proprietary format model is used for model transfer/delivery, delta model update definition copied from legacy delta signaling definition is impossible as the gNB usually cannot recognize the details of the AI/ML model algorithm for a proprietary format model.  More addition, even if we consider open format model for solution1a, the model privacy and future proof are still under estimation. Every time a new open format model is introduced into 3GPP system, all legacy gNBs should be upgraded to understand the new model, which is somehow impossible/undesirable from operator perspective, so how open format model can work for future proof is still questionable for solution1a.  **Observation: Delta model update definition ported from legacy delta signaling definition is only applied to open format case for solution1a, but not applied to proprietary-format case.**  Based on above, we propose:  **Current status and Gaps:**  May be supported if open format is used for model transfer  **RAN specification impact**  How partial model update is applied to proprietary format may need extra spec effort.  A8: See comments in Q1 |
|  |  |  |
|  |  |  |

**Q3-1a: For Solution 1a, is there any other readiness/current status and gaps/RAN specification impact need to be added in the above table? Please state from the objective fact point of view.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Readiness** | **Current status and gaps** | **RAN specification impact** |
| #example | Ax: | Ax: | Ax: |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

##### Solution 2a/3a: CN (except LMF)/LMF can transfer/deliver AI/ML model(s) to UE via NAS signalling/LPP signalling

Table . Solution 2a/3a Readiness and RAN specification impact

|  |  |  |
| --- | --- | --- |
| **Discussion Area** | **Readiness** | **RAN specification impact** |
| **Current status and Gaps** |
| A1 | not supported ☹ | extension of the number of RRC segments is required to support models larger than 45kBytes |
| model size >45kBytes is not supported based on existing number of RRC segments |
| A2 | supported 😊 |  |
| A3 | 1) procedure latency depends on model size and SRB priority; 2) other latency includes forwarding NAS message latency from CN to gNB |  |
| A4 | supported with limitation 😊 |  |
| For Solution 2a, support within AMF coverage area based on PDCP status report;  For Solution 3a, support within LMF coverage area based on LPP signaling segmentation |
| A5 | not supported ☹ | support management and model transfer interaction between CN and gNB |
| For Solution 2a, gNB cannot perform management directly, considering model transfer is transparent to gNB |
| A6 | Solution 2a: not supported ☹  Solution 2b: supported 😊 | For solution 2a, support delta signaling for NAS message |
| Solution 2a: NAS delta signaling is not supported |
| A7 | not supported ☹ | introduce multiple SRBs or SRB with variable/multiple priorities |
| SRB priority is used |
| A8 | supported 😊 |  |

**Q2-2a3a: For Solution 2a/3a, do you agree the content in above table to capture analysis of model transfer/delivery solutions? (please include comments only on the current context in this question. For new add-ons, please see Q3-2a3a)**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comment** |
| #example | Yes: Ax No: Ay | Ax:  Ay: |
| OPPO | No for A1/A2/A3/A4/A5/A6/A7/A8 | A1: Upper layer segmentation is invisible for RRC layer, we propose the following rewording:  **Current status and Gaps:**  Although upper layer segmentation is invisible for RRC layer, SRB1 or SRB2 will be used to carry the continuous upper layer segmentation container, this may have influence on transmission of other information, e.g. normal NAS/RRC info, carried via SRB1/SRB2.**RAN specification impact**  ]  The coexistence between upper layer segmentation container including model data, and normal NAS/RRC info.A2: See comments in Q1  A3: RAN specification impact is missing, we propose:  **RAN specification impact**  If SRB other than SRB1/SRB2 is introduced or reused, RAN may need to consider the SRB priority.  A4: Upper layer segmentation is invisible for RRC layer, we propose the following rewording:  **Current status and Gaps:**  For Solution 2a, support based on NAS signaling segmentation;  For Solution 3a, support based on LPP signaling segmentation.  **RAN specification impact**  The coexistence between upper layer segmentation container including model data, and normal LMF triggered/NAS/RRC info.  A5:  **Current status and Gaps:**  For Solution 2a/3a, gNB cannot perform management directly, considering model transfer is transparent to gNB  **RAN specification impact**  For solution 2a, support management and model transfer interaction between CN except LMF and gNB via NAS signaling;  For solution 3a, support management and model transfer interaction between LMF and gNB via NRPPa signaling;  A6  We don’t think RAN can evaluate the delta configuration for upper layer, so a general description is sufficient and safe:  **Current status and Gaps:**  Upper layer delta configuration is invisible from RAN point of view.  **RAN specification impact**  Note: delta configuration may have some spec impact for CN.  A7: Currently, QoS requirements are not applicable to SRB as SRB priority is used instead.  **Current status and Gaps:**  N/A  **RAN specification impact**  N/A  A8: See comments in Q1 |
|  |  |  |
|  |  |  |

**Q3-2a3a: For Solution 2a/3a, is there any other readiness/current status and gaps/RAN specification impact need to be added in the above table? Please state from the objective fact point of view.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Readiness** | **Current status and gaps** | **RAN specification impact** |
| #example | Ax: | Ax: | Ax: |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

##### Solution 1b: gNB can transfer/deliver AI/ML model(s) to UE via UP data

Table . Solution 1b Readiness and RAN specification impact

|  |  |  |
| --- | --- | --- |
| **Discussion Area** | **Readiness** | **RAN specification impact** |
| **Current status and Gaps** |
| A1 | supported 😊 | support DRB termination at gNB |
| A2 | supported? |  |
| A3 | procedure latency depends on model size and DRB priority |  |
| A4 | not supported ☹ | identify a solution to support service continuity support between gNBs when DRB is terminated at gNB |
| Solution not identified to support model transfer continuity if DRB terminated at gNB |
| A5 | supported 😊 |  |
| A6 | not supported ☹ | solution for gNB to support delta-model transfer/delivery in user plane |
| solution not identified to support model update if DRB terminated at gNB |
| A7 | supported 😊 |  |
| A8 | supported 😊 |  |

**Q2-1b: For Solution 1b, do you agree the content in above table to capture analysis of model transfer/delivery solutions? (please include comments only on the current context in this question. For new add-ons, please see Q3-1b)**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comment** |
| #example | Yes: Ax No: Ay | Ax:  Ay: |
| OPPO | No for A1/A2/A3/A7/A8 | A1: Only ‘support’ is not accurate to reflect the status, so we propose:  **Current status and Gaps:**  No model size limitation if UP method is used for model transfer  A2: See comments in Q1  A3: QoS requirements should be considered also for DRB, so we propose:  **Current status and Gaps:**  procedure latency depends on model size, QoS requirements and DRB priority  **RAN specification impact**  QoS management for model transfer  Note: Whether QoS management for solution1b has CN involvement needs SA clarification.  A7: In legacy, gNB is not the entity to control QoS, so we propose:  **RAN specification impact**  QoS management for model transfer  Note: Whether QoS management for solution1b has CN involvement needs SA clarification.  A8: See comments in Q1 |
|  |  |  |
|  |  |  |

**Q3-1b: For Solution 1b, is there any other readiness/current status and gaps/RAN specification impact need to be added in the above table? Please state from the objective fact point of view.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Readiness** | **Current status and gaps** | **RAN specification impact** |
| #example | Ax: | Ax: | Ax: |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

##### Solution 2b/3b: CN (except LMF)/LMF can transfer/deliver AI/ML model(s) to UE via UP data

Table . Solution 2b/3b Readiness and RAN specification impact

|  |  |  |
| --- | --- | --- |
| **Discussion Area** | **Readiness** | **RAN specification impact** |
| **Current status and Gaps** |
| A1 | supported 😊 | No RAN impact  Note: The detail procedure of model transfer from CN/LMF to UE is out of RAN scope |
| A2 | supported 😊 |  |
| A3 | 1) procedure latency depends on model size and DRB priority; 2) other latency includes forwarding data from CN to gNB |  |
| A4 | support with limitation 😊 |  |
| For Solution 2a, support within AMF coverage area based on PDCP status report; For Solution 3a, support within LMF coverage area based on LPP signaling segmentation |
| A5 | gNB cannot perform model management directly, NAS signalling is used to configure and initiate model transfer from CN. | support management and model transfer interaction between CN and gNB |
| A6 | not supported ☹ | solution for CN to support delta-model transfer/delivery in user plane |
| CN cannot support delta-model transfer/delivery in user plane |
| A7 | supported 😊 |  |
| A8 | supported 😊 |  |

**Q2-2b3b: For Solution 2b/3b, do you agree the content in above table to capture analysis of model transfer/delivery solutions? (please include comments only on the current context in this question. For new add-ons, please see Q3-2b3b)**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comment** |
| #example | Yes: Ax No: Ay | Ax:  Ay: |
| OPPO | No for A1/A2/A3/A4/A5/A7/A8 | A1: Although the content of GTP/UP tunnel may be transparent to gNB, the gNB may still need to get the model meta info from CN/LMF, this may fall into the A5 scope, so we propose:  **Current status and Gaps:**  No model size limitation if UP method is used for model transfer  **RAN specification impact**  See the impact analysis for A5A2: See comments in Q1  A3: QoS requirements should be considered also for DRB, so we propose:  **Current status and Gaps:**  1) procedure latency depends on model size , QoS requirements and DRB priority; 2) other latency includes forwarding data from CN to gNB  **RAN specification impact**  QoS management for model transfer  Note: Whether QoS management for solution1b has CN involvement needs SA clarification.  A4: For solution 2b, we don’t know why we only focus on AMF coverage case, we think NG HO case can also result in AMF change and Model transfer/delivery continuity is also guaranteed via UP L2 handling. More addition, PDCP status report is a specific solution, we should make the wording generic during SID phase.  For solution3b, we understand the proposed wording is a typo, because it’s obvious that solution3b does not rely on LPP signaling segmentation. Maybe we can merge solution2b/3b, so we propose:  **Current status and Gaps:**  For Solution 2b/3b/ Model transfer/delivery continuity is achieved via UP L2 handling, e.g. PDCP status report.  A5: Better to differentiate solution2b/3b  **Current status and Gaps:**  For solution2b/3b, gNB may need extra method to acquire model meta info for model management purpose;  **RAN specification impact**  For solution2b, gNB cannot perform model management directly, NAS signalling is used to configure and initiate model transfer from CN;  For solution3b, gNB cannot perform model management directly, NRPPa signalling is used to configure and initiate model transfer from LMF.  Note: whether RAN3/SA2 is involved may need RAN3/SA2 clarification.  A7: new QoS policy may be considered by CN, but the details should be clarified by SA2.  **RAN specification impact**  Note: whether SA2 is involved may need SA2 clarification  A8: See comments in Q1 |
|  |  |  |
|  |  |  |

**Q3-2b3b: For Solution 2b/3b, is there any other readiness/current status and gaps/RAN specification impact need to be added in the above table? Please state from the objective fact point of view.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Readiness** | **Current status and gaps** | **RAN specification impact** |
| #example | Ax: | Ax: | Ax: |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

##### Solution 4a: OTT server can transfer/delivery AI/ML model(s) to UE (transparent to 3GPP)

Table . Solution 4a Readiness and RAN specification impact

|  |  |  |
| --- | --- | --- |
| **Discussion Area** | **Readiness** | **RAN specification impact** |
| **Current status and Gaps** |
| A1 | supported 😊 | No RAN impact |
| A2 | Not within RAN scope |
| A3 | 1) procedure latency depends on model size and DRB priority; 2) other latency includes forwarding data from OTT server to gNB |
| A4 | not supported ☹ |
| transparent to RAN |
| A5 | not supported ☹ |
| transparent to RAN |
| A6 | not supported ☹ |
| transparent to RAN |
| A7 | supported 😊 |
| A8 | not supported ☹ |

**Q2-4a: For Solution 4a, do you agree the content in above table to capture analysis of model transfer/delivery solutions? (please include comments only on the current context in this question. For new add-ons, please see Q3-4a)**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comment** |
| #example | Yes: Ax No: Ay | Ax:  Ay: |
| OPPO | No for A2/A3/A4/A6/A8 | A2: See comments in Q1  A3: QoS requirements should be considered also for DRB, we also don’t know how OTT server can transfer the model to gNB directly, so we propose:  Current status and Gaps:  1) procedure latency depends on model size , QoS requirements and DRB priority; 2) other latency includes forwarding data from OTT server to CN  RAN specification impact  QoS management for model transfer  Note: Whether QoS management for solution4a has CN involvement needs SA clarification.  A4: solution4a is also a special kind of UP solution, Model transfer/delivery continuity can also be achieved via UP L2 handling, so we propose:  **Current status and Gaps**  May besupported via UP L2 handling.  A6: transparent to RAN is sufficient, we cannot say there is no Partial model update via Non-3GPP method, so we propose:  **Current status and Gaps**  transparent to RAN  A8: See comments in Q1 |
|  |  |  |
|  |  |  |

**Q3-4a: For Solution 4a, is there any other readiness/current status and gaps/RAN specification impact need to be added in the above table? Please state from the objective fact point of view.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Readiness** | **Current status and gaps** | **RAN specification impact** |
| #example | Ax: | Ax: | Ax: |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

##### Solution 4b: OAM can transfer/delivery AI/ML model(s) to UE

Table . Solution 4b Readiness and RAN specification impact

|  |  |  |
| --- | --- | --- |
| **Discussion Area** | **Readiness** | **RAN specification impact** |
| **Current status and Gaps** |
| A1 | over CP: not supported ☹  over UP: supported 😊 | over CP: if model size larger than 45kBytes, extend RRC segment number |
| over CP: maximum 45kBytes based on existing number of RRC segments |
| A2 | Not within RAN scope | Note: The details security and integrity of solution 4b is out of RAN scope |
| A3 | 1) procedure latency depends on model size and SRB/DRB priority; 2) other latency includes forwarding data from OAM to gNB | latency reduction if model transfer/delivery has critical/relax latency requirement |
| A4 | support within OAM coverage 😊 |  |
| A5 | not supported ☹ | support management and model transfer interaction between OAM and gNB |
| gNB cannot perform model management directly, signalling between gNB and OAM is used to configure and initiate model transfer from OAM. |
| A6 | over CP: not supported ☹  over UP: not supported ☹ | support delta signaling/delta-model transfer/delivery over CP/UP |
| over CP: OAM delta signaling to gNB is not supported  over UP: user plane cannot support delta-model transfer/delivery |
| A7 | supported 😊 |  |
| A8 | not supported ☹ |  |

**Q2-4b: For Solution 4b, do you agree the content in above table to capture analysis of model transfer/delivery solutions? (please include comments only on the current context in this question. For new add-ons, please see Q3-4b)**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comment** |
| #example | Yes: Ax No: Ay | Ax:  Ay: |
| OPPO | Not sure for A1  No for A2/A3/A8 | A1: for solution4b, we are wondering why the segmentation is visible from RRC point of view if OAM control the segmentation, in basic logic, RRC layer should treat any OAM segmentation as a container like any other upper layer info.  A2: See comments in Q1  A3: QoS requirements should be considered also for DRB  **Current status and Gaps**  1) procedure latency depends on model size , QoS requirements(DRB only) and SRB/DRB priority; 2) other latency includes forwarding data from OAM to gNB  A8: See comments in Q1 |
|  |  |  |
|  |  |  |

**Q3-4b: For Solution 4b, is there any other readiness/current status and gaps/RAN specification impact need to be added in the above table? Please state from the objective fact point of view.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Readiness** | **Current status and gaps** | **RAN specification impact** |
| #example | Ax: | Ax: | Ax: |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Conclusion

TBD

# 4. References

[1] R2-2302268 Report of Offline 027 model transfer delivery (Huawei)

[2] R2-2308286 Report of [Post122][060][AIML] Mapping of functions to physical entities (CMCC)