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)

**A9: Deployment/enhancements to network interfaces**

**A10: gNB complexity (e.g., storage and processing)**

**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.** |
| Huawei, HiSilicon | No: A3, A7, A6  Yes: Others | **A3, A7**  We think A3 and A7 can be merged as both are about QoS impacts.  **A6**  For the terminology partial model update, the meaning is unclear and what RAN2 should study is also unclear.  RAN1 has the definition of model update:  **Retraining or fine tuning of an AI/ML model, via online/offline training, to improve the model inference performance.**  However, there were not much progress on details.  In previous RAN2 discussions, some companies thought that if some of model information is visible to some layer, delta configuration may be used. However, this was not discussed and confirmed.  In general, we do not think the necessity of A6 is clear.  **Others**  We observe that RAN1 is also discussing case y/z1-z5 for model transfer/delivery. For pros/cons analysis in RAN1, they are discussing the evaluation metrics, and maybe some of metrics are similar to what we are discussing here.  For now, we think RAN2 and RAN1 can have parallel discussions/analysis. |
| Qualcomm | No for A2, A3, A5, A6, and A8  Please add A9 and A10 | A2  Whether model delivery is over control or user plane, legacy procedures already support security aspects.  A3  It is not realistic to ask every gNB to store all models. A central storage for gNB-based solutions will make delays for all solutions similar (storage -> gNB -> UE).  A5  For all of the model delivery, the LCM decisions can still remain at the gNB. For example, in model ID-based LCM, even in 1b/2a/2b/4, which model should be used can be determined by the gNB. Therefore, in all model delivery methods considered, the network controls the LCM.  A6  The user plane can support delta model delivery. It is up to the implementation, how that model and parameter sets are stored. Model training entities can develop multiple parameter sets for a model structure. These parameter sets can be transferred using UP when required. Therefore, delta model delivery can be supported in all model delivery methods.  A8  For all of the model delivery methods, inter-operability issues exist. For example, if the model is developed by a UE vendor, then gNB still needs to know how to manage the model at the UE (for model ID-based LCM). Therefore, all solution has the same inter-operability.  Furthermore, the inter-operability issues remain valid even for open-format models, as the entities that are developing the open-format models or parameter sets still have to indicate to the NW how these open-format models are to be used. If developed by the network, then, cross-compiling UE models is an issue that is not solved.  **We also propose to add the following issues:**  A9: Deployment Impact  Some solutions have greater deployment impact than others. The differences in deployment impact should be studied.  A10: gNB impact (e.g., standard interface, storage and processing)  Some solutions have greater gNB impact than others for standardization and implementation. The differences in gNB impact should be studied. |
| Apple | OK to discuss all (A1-A10) in this email discussion to identify potential spec impacts.  But disagree to capture any of them in TR 38.843 as "requirement" or "readiness" of model transfer. | Thanks for Rapporteur's hard work. We are fine to **consider A1-A10** **only in this email discussion to identify potential spec impacts if converged** (e.g. only capture "increase segmentation number" for solution 1a if we can converge this point, but no need to capture its readiness column).  However, we **do not agree to capture any of them in TR 38.843 as "requirement" or "readiness" of model transfer.** Our considerations are:   1. RAN2 is only responsible for a small piece of model transfer (i.e. signaling). It is still RAN1 to determine requirement, feasibility and conclusion of model transfer. 2. Among A1-A10, some of them are just enhancement direction while some of them are not clear whether they are requirement:    * For example, A1 may not be a requirement if RAN1 conclude that model for Rel-18 use cases is expected to be smaller than 45kbyte.    * For example, A3/A7 are QoS rather than requirements (i.e. workable or not).    * For example , A4/A6 are actually enhancement because whether to support service continuity and delta signaling doesn't impact whether model transfer can work but just performance enhancement. 3. We have agreed the table with Pros and Cons in RAN2#121. The agreed Pros and Cons + potential RAN impacts in this email discussion are sufficient to conclude model transfer in SI from RAN2 perspective. From this point of view, extra capturing "Readiness" or "Requirement" is not necessary.   Thus, if any of them are captured in TR, it means RAN2 have to satisfy it. We don't think RAN2 is ready to make such conclusion. So, we don't agree to capture anything in "**readiness**" column of followed 8 tables (i.e. we agree to capture " **RAN specification impact** " column if converged). |
| ZTE | Yes: A1, A4, A6  No: A2, A3, A5, A8, A9, A10  Neutral: A7 | **A1.** **Large, no upper limit model size** (mentioned in Solution 1a, Solution 2a and 3a, Solution 1b, Solution 4)  **ZTE: The data transmission have a upper boundary via CP tunnel while there is no upper boundary via UP tunnel, it can be one of the benchmarks to evaluate the PRO and CONs for each solution**  **A2. Security and integrity** (mentioned in Solution 1a)  **ZTE: In our understanding, not only CP based solution, both UP transmission also can be ciphered/IP which depends on the RRC configuration. This evaluation can not be a benchmark for evaluating the solutions**  **A3. Latency requirement, e.g. critical, relax, no latency requirement** (mentioned in Solution 2a)  **ZTE: It is really confusing, The latency requirement seems a kind of requirement evaluation for different use cases of model transfer rather than an evaluation for the model transfer solutions. In addition, since we does not do enough research for the solutions on the table, the delay evaluation for each solution may not be scientific/realistic.**  **A4. Model transfer/delivery continuity (i.e. resume transmission of model (segments) across gNBs)** (mentioned in Solution 1a, Solution 2a, Solution 1b)  **ZTE: Yes, CP based solution will encounter such issue.**  **A5. NW controllability (e.g. model management decision at gNB)** (mentioned in Solution 1a, Solution 2a)  **ZTE: In our understanding, the NW controllability cannot be a benchmark for the evaluation of solution. It is hard to say the controllable model transfer is a good thing (PROs), otherwise is not (CONs)...**  **A6. Partial model update (e.g. delta configuration)** (mentioned in Solution 1a, Solution 2b and 3b)  **ZTE: Yes, if the open format can be supported, the partial model update can be realized in the CP based solution which can save the overhead of model transfer.**  **A7. Flexible model transfer/delivery QoS** (mentioned in Solution 1b, Solution 2b and 3b, Solution 4)  **ZTE: Not sure whether there is any need to classify the different model transfer with different QoS. We can follow the majorities.**  **A8. Interoperability (e.g. No/minor need for offline coordination among vendors)** (mentioned in Solution 4)  **ZTE: In our understanding, all model transfer between UE and NW may have interoperability issue regardless of the which solution is used.**  **A9: Deployment/enhancements to network interfaces**  **ZTE: the Network interface is out of RAN2 spec which has RAN3/SA impact.suggest not using this as benchmark to evaluate the solution.**  **A10: gNB complexity (e.g., storage and processing)**  **ZTE: It is not in the 3GPP scope which cannot be listed in the 3GPP TR.** |
| Mediatek | No: A2  Yes: A1, A3(with revision), A4, A7, A5 (with revision),  Need clarification on **A6, A8,** A9, A10 | **A2**  Same view as OPPO and Qualcomm.  **A3**  Based on the description, it would be more accurate to refer to it as '**the overall latency of model transfer/delivery**' rather than 'latency requirement',  as the latency requirement (critical, relax, or no latency requirement) comes from RAN1 on how quickly model transfer/delivery needs to be completed, e.g., for reactive model transfer/delivery.  **A5-> suggested to revised as** NW controllability on model transfer/delivery  The intention is NW controllability on model transfer/delivery instead of other LCM aspects, e.g. model management. For all solutions, NW controllability on model management (model monitoring, model activation/deactivation/switch/fallback) is the same.  **A6**  The specifics of how a partial model update operates and how a delta configuration could be possible are not clear. This point may be more closely related to the model format, i.e., whether it's an open format or a proprietary format. A partial model update typically involves updating parameters, rather than the entire model. The feasibility of a delta configuration, on the other hand, could depend on various factors, e.g. how to representing the model structure and parameters with ASN.1. RAN2 needs to clarify how it works first.  **A7**  Need to clarify the motivation to differentiate the model transfers with different QoS/SRB priorities, and whether flexible QoS is needed for model delivery/transfer.  **A8**  The exact meaning of interoperability and the types of offline coordination considered are unclear. This point may also be related to the format of the model, i.e., whether it's an open format or a proprietary format. RAN2 needs to clarify what interoperatibility issues are concerned.  One valid point may be whether model/functionality identification is performed in offline manner or through signlaing over air interface. If it is the only point, we may need to make it this point clear.  **A9: Deployment/enhancements to network interfaces**  I'm uncertain whether we need to consider the realistic deployment status for this discussion, as there may be more constraints in a real network. The complexity and variability of real-world network deployments could introduce additional factors that we need to consider. Therefore, we need to understand the implications of these factors on our discussion.  **A10: gNB complexity (e.g., storage and processing)**  Why only gNB complexity is concerned instead of the NW complexity? It might be due to the specific role and function of gNBs in the network. However, the complexity can vary depending on where the model is stored. This is particularly relevant when considering different location assumptions. |
| Interdigital | See comments | We agree with the sentiment expressed by Apple that the discussion points are just to facilitate this email discussion and not to be captured as is in the TR.  We also think it will facilitate the discussion if we remove some items.  *To remove*  *A2*: As expressed by other companies as well, we can remove this one as both CP/UP support both encryption and integrity protection.  *A6:* It is not clear if model update is something that happens that often to make the delta update an important aspect to consider. As also pointed out by other companies, even for the UP solution, delta update can be made (e.g., if open format is used)  *A8:* As others have pointed out, there could be interoperability issues in most of the cases  *A9/A10*: We agree with the views expressed by ZTE regarding these points.  Thus, in all the questions below, we are providing comments only on A1, A3, A4, A5 and A7 |
| Xiaomi | No: A2, A6  Yes: A1, A3, A4, A5, A7  New area (model delivery in uplink) should be also considered. | **A2:** Agree with others that in all solutions, ciphering and integrity protection are applicable, therefore no need to list it as aspect for comparison.  **A6**: Agree with others that it is not clear that delta update is an important area to consider, and UP solutions can also support such delta update when open format is used.  New aspect: we think **model delivery in uplink** should be supported since RAN2 agreed to support model delivery from UE to gNB for type 1 of CSI compression with two-sided model [2]. Therefore it should be listed as a new area. |
| TCL | No: A2,  Yes with comments: A6, A3, A7,  Yes: Others. | A2:  We show similar views with some of above companies, in the legacy methods including CP and UP solutions, the security and integrity are already guaranteed.  A6:  It is not clear to us what is partial model update (e.g. delta configuration)? Suggest to first clarify the definition of the partial model update.  In our understanding, partial model update means that some parameters have been redeveloped for a certain model structure, such as learnable parameters, hyper-parameters. Firstly, from our perspective, the parameters of partial model update may be regraded as data packet, which can be transferred/delivered through CP and UP solutions, but for proprietary format model, the UE vendor can not recognize the parameters from NW vendor, it means that the partial model update transfer/delivery is meaningless. On the other hand, with open format model, the above parameters of a model are visible for two different vendors, only the changed parameters can be transferred/delivered, which is able to decrease the signaling overhead of using CP solutions compared to UP solutions.  A3 and A7:  The model transfer/delivery may have various latency requirements, which may affect the priority of SRB/DRB and QoS. Given there is the mapping between SRB/DRB and QoS, we think it is better to merge the discussion about A3 and A7. |

## 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  =>A9: Requires Xn and/or NG-AP Interfaces.  =>A10: Requires gNB to store models. If not stored locally, then, latency is increased as well. |
| **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.  =>A10: No gNB Impact. | 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.  =>A9: Additional deployment impact out of RAN2 scope. |
| **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  =>A10: Requires gNB to store models. If not stored locally, then, latency is increased as well. |
| **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  =>A10: No additional gNB impact. | 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  =>A9: Additional deployment impact out of RAN2 scope. |
| **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  =>A9: No additional deployment impact.  =>A10: No additional gNB impact. | 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 |  |
| A9 | Requires Xn and/or NG-AP Interfaces |  |
| A10 | gNB complexity (storage and processing) |  |

**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 |
| Huawei, HiSilicon | No: A3, A7, A6  Yes: Others | **A3, A7**  As we commented above, both A3 and A7 are about QoS impacts, so there could be the uniform analysis.  For readiness of A3 and A7, the text is suggested: **Not supported. It depends on model size and SRB priority.**  For RAN spec impact, the text is suggested: **Impacts on SRBs in DL, e.g. introduce multiple SRBs or SRB with variable/multiple priorities.**  **A6**  As we commented above, the meaning of A6 is unclear and it should be clarified first. e.g. what kind of information may need delta configuration. |
| Qualcomm | Yes, for all, with comments.  A2, A3, A5, A6, A8 are similar for all solutions, therefore A2, A3, A5, A6, and A8 should be removed from the table. Instead, A9, and A10 should be added. | A2: See comments to Q1. Remove 9 (in pros) from solution 1a in the pros and cons table.  A3: Either every gNB will have to store all models (see A10), or a centrally located storage will have similar delays as any other solution.  A5: As discussed in response to Q1, in all of the solutions, NW (gNB) can manage the model. Note that the model can be developed by UE vendors or 3rd party. For solution 1a, we still need a solution for model transfer/delivery from the training entity to gNB. Remove 6 and 11 (in pros) from solution 1a in the pros and cons table.  A6: discussion on A6 for solution 1a is irrelevant. Also, see the comment to Q1. All the model transfer/delivery methods can support parameter set update/delta model update. This is up to model development.  A8: See comment to Q1.  A9: Currently, Xn is not widely deployed in the network. For this delivery method to work, we need high Xn deployment.  A10: Needs more storage and processing at gNB. If not stored locally, then, latency is increased as well. |
| Apple | Yes: A1, A4 with comments, A7  No: all others | On A4, please note that current SRB4 (QoE) has already support service continuity during HO. It is just segmentation transmission continuity can't be ensured because UE will discard the previous transmitted RRC segments during HO, according to section 5.3.5.3 of TS 38.331:  *3> if configured with application layer measurements and if application layer measurement report container has been received from upper layers for which the successful transmission of the message or at least one segment of the message has not been confirmed by lower layers:*  *4>* ***re-submit the MeasurementReportAppLayer message or all segments of the*** *MeasurementReportAppLayer message to lower layers for transmission via SRB4;*  Thus, we suggest to revise A4 as:  Introduce service continuity support for SRBs with segmentations. |
| ZTE | Yes : A1, A4, A6  No with comments: A7  No: All others | We tend to agree with the A1, A4, A6 which is summarized by rapporteur.  Regarding A7, it is not crystal clear about the motivation to differentiate the model transfers with different QoS/SRB priorities. We need to confirm the motivation first, and then to discuss whether the multiple SRBs for model transfer is needed or not.  For other items except for A1, A4, A6, please see our reply in question 1 where all other items seems not be precisely/correctly to reflect the PROs and CONs for each solution. |
| Mediatek | Yes: A1, A3 with change, A4, A5 with change  FFS: A6, A7, A8, A9, A10 | A3 yes, if the overall latency is concerned (instead of latency requirement)  A5 yes, if the NW controllability focuses on the controllability on model transfer/delivery.  A6, A7, A8 needs further clarification  A9 see comment in Q1  A10  True for solution 1a, but not clear about the specification impact |
| Interdigital | No (see comment to Q1): A2, A6, A8, A9, A10  No: A3 (see comments)  Yes: A1, A4, A5, A7 | A3: large model size support is already discussed under A1, so assuming this is addressed, it seems that solution 1a will not have latency issues? |
| Xiaomi | Comments for A3  No need to discuss A2 and A6 (see comments to Q1)  Yes for others | A3: Solution 1a has impact on latency of other RRC messages since current requirement in TS 38.300 is that *“all segments of an RRC message are transmitted before sending another RRC message*”. This means that when transmitting a large AI model, other RRC messages, which could be delay sensitive, might be delayed. |
| TCL | No: A2,  Yes with comments: A6, A3, A7,  Yes: Others | A2: See comments in Q1.  A6: See comments in Q1.  A3 and A7:  There is a little bit of repeat and conflict on the discussion of RAN specification impact for A3 and A7. As described in the Q1’s comments, it is possible that model transfer/delivery may have different latency requirements, which may bring some spec impact on radio bearer priority setting, we suggest to marge A3 and A7 and align the discussion. |

**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: |
| Qualcomm | Can potentially support transfer or delivery of models < 45KBs   * Even in that case the model delivery can be interrupted (fail) due to handover | Significant gap   * Delivery of large model * Lossless model delivery * Dependence on availability of Xn interface   If Xn is not deployed between two gNB, then any proposed enhancements cannot work. | Requires RAN2 enhancements such as   * Larger RRC segmentation * SRB Reestablishment, such that model delivery can continue.   Requires Xn/NG-AP enhancements such as   * Forwards of untransmitted model or sequence number   This method highly depends on Xn and/or NG-AP enhancements/deployments. Otherwise, none of the RAN2 proposed solutions work. |
| Xiaomi |  | Model delivery in uplink | New RRC message and procedure are required to support AI model delivery in uplink for solution 1a since existing messages like UAI is not suitable to deliver AI model. |
|  |  |  |  |

##### 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 |  |
| A9 | Impact out of RAN2 scope |  |
| A10 | No additional gNB impact |  |

**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 |
| Huawei, HiSilicon | No: A1, A3, A7, A6  Yes: Others | **A1**  For segmentation of CP signaling, it depends on the CP level. RAN2 did not discuss in which layer the segementation should be performed, and we think at least RRC or NAS/LPP may be candidate entities for doing the segmentation. If RRC does the segmentation, there could be the similar analysis to Solution 1a; if NAS/LPP does the segmentation, it may only impact NAS layer or LPP layer, and thus the impacts on RRC may be minimal.  So we suggest:  For readiness, ok with the current wording.  For RAN spec impact, change the text into:  **If NAS/LPP do not support segmentation, it may need RRC segmentation, and extension of the number of RRC segments is required to support models larger than 45kBytes.**  **A3, A7**  As we commented above, both A3 and A7 are about QoS impacts, so there could be the uniform analysis.  For readiness of A3 and A7, the text is suggested: **Not supported. It depends on model size and SRB priority. Other latency includes forwarding NAS message latency from CN to gNB.**  For RAN spec impact, the text is suggested: **Impacts on SRBs in DL, e.g. introduce multiple SRBs or SRB with variable/multiple priorities.**  **A6**  As we commented above, the meaning of A6 is unclear and it should be clarified first. e.g. what kind of information may need delta configuration. |
| Qualcomm | No for A1, A4, A5, A6  A2, A3, A5, A6, A8 are similar for all solutions, therefore A2, A3, A5, A6, and A8 should be removed. Instead, A9 and A10 should be added. | A1: NAS can segment in such a way that RRC segmentation may not be required.  A2: See comments to Q1.  A5: Same as comments to Q1, Q2-1a. Remove 4 (in cons) from solution 2a/3a in the pros and cons table.  A6: Same as comments to Q1, Q2-1a. |
| Apple | Yes: A1 with change  No: all others | We only agree A1 with Huawei's change, i.e. **If NAS/LPP do not support segmentation, it may need RRC segmentation, and extension of the number of RRC segments is required to support models larger than 45kBytes.**  For A5/A6, we think their spec impact are out of RAN2 scope. So, RAN2 is not in position to make conclusion (i.e. not capture them in TR).  On A7, it essentially requires a new QoS profile for model transfer in UP. It is SA2 scope. So, RAN2 is not in position to make conclusion (i.e. not capture them in TR). |
| ZTE | Yes for readiness: A1, A4, A6,  Yes for specification impact: A6  Yes for specification impact with modification: A1.  No for A7.  No for All others. | Regarding the specification impact for A1, we share the same view with HW/apple regarding solution 2a. For the solution 3A already support segmentation (i.e. LPP), **we do not think there is any enhancement to RRC segmentation needed for solution 3A.** Regarding solution 2A, potential specification impact in both SA and RAN can be clarified, In this sense, we suggest to modify the specification impact for A1 like below:  NAS signaling segmentation or extansion of RRC segmentation maybe required for solution 2A.Regarding A7, please see our comments in Question 1 and 2.  For other items except for A1, A4, A6, please see our reply in question 1 where all other items seems not be precisely/correctly to reflect the PRO and CONs for each solution. |
| Mediatek | Yes: A1 with change, A3, A4 with change, A5,  FFS: A6, A7, A8, A9, A10 | A1:  We agree with HW’s revision.  A4:  For Solution 2a, support within AMF coverage area based on NAS signaling segmentation.  For Solution 3a, support within LMF coverage area based on LPP signaling segmentation  A6, A7, A8: see comment in Q1  A9 see comment in Q1. If it is out of RAN2 scope, why we need to list it in the table?  A10 there should be some impacts on gNB e.g., due to the enhancement on NRPPa signaling. |
| Interdigital | No (see comment to Q1): A2, A6, A8, A9, A10  No: A1 (see comments), Yes: A3, A4, A5, A7 | A1: Agree with the comments from Huawei/Apple |
| Xiaomi | No need to discuss A2 and A6 (see comments to Q1)  Yes for others |  |
| TCL | No: A1, A2, A6;  Yes with comments: A6, A3, A7;  Yes: Others. | A1: Agree with Huawei's views, both RRC or NAS/LPP are able to perform the segmentation.  A2: See comments in Q1.  A6: See comments in Q1.  A3 and A7: See comments to Q1, Q2-1a. |

**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: |
| Qualcomm | DL NAS transfer message can be reused. | Significant gap   * Delivery of large model * NAS or RRC segmentation can needs enhancements. * Lossless model delivery   Indication of model download completion. | Some signalling enhancements may be required to indicate  Model transfer is completed. |
|  |  |  |  |
|  |  |  |  |

##### 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 |  |
| A9 | Requires Xn and/or NG-AP Interfaces |  |
| A10 | gNB complexity (storage and processing) |  |

**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 |
| Huawei, HiSilicon | No: A2, A3, A7, A6  Yes: Others | **A2**  Readiness should be “**not supported**” as solutions are unclear for now.  RAN spec impact: **support DRB termination at gNB**  **A3, A7**  As we commented for above questions, there should be uniform analysis for A3 and A7 as they are about QoS impacts.  **A6**  As we commented for above questions, the meaning of A6 is unclear, and what kind of information may need A6 is also unclear. |
| Qualcomm | No for A2, A4, A6  A2, A3, A5, A6, A8 are similar for all solutions, therefore A2, A3, A5, A6, and A8 should be removed. Instead, A9 and A10 should be added. | A2: See comments to Q1.  A4: RAN2 has no expertise on this. I-UPF at two gNB can be connected and transfer the status of model transfer/delivery between themselves. Remove 5 (in cons) from solution 1b in the pros and cons table.  A6: See comments to Q1 and Q2-1a. The assumption that delta or parameter set update cannot be supported over the UP-based method is wrong. |
| Apple | Yes: A1 with change, A4  No: all others | We agree with Huawei that for A1:   * Readiness should be “**not supported**” as solutions are unclear for now. * RAN spec impact: **support DRB termination at gNB**   For A6, we may understand Rapporteur intention, but we prefer to avoid confusing term like "delta signaling in UP" which should not be captured in TR. |
| ZTE | Yes for readiness: A1, A4, A6  No for specification impact: A4  Yes for specification impact: A1, A6  No for all others | Regarding the specification impact for A4, in our understanding, the data continuity is mainly for HO case, and data forwarding is already supported for inter-CU handover, so we do not think there is any specification impact even if the DRB is assumed to be terminated at gNB.  Regarding A7, please see above comments from us  For other items except for A1, A4, A6, please see our reply in question 1 where all other items seems not be precisely/correctly to reflect the PRO and CONs for each solution. |
| Mediatek | No: A2  Yes: A1 with change  FFS A6, A8, A9, A10  The assumption on A3, A4. A5, A7 need to be clarified. | A1: Agree with OPPO  A3, A4, A5 and A7 in the table 3 take the assumption that the DRB carrying model is terminated at the gNB, which has direct control on QoS.  But the alternative way is that the DRB carrying the model is terminated at UPF as existing DRBs, which requires the gNB delivers the model to the UPF. Then we may have different conclusion.  A6, A7, A8 see comment in Q1 |
| Interdigital | No (see comment to Q1): A2, A6, A8, A9, A10  No: A3, A4 (comments)  Yes: A1, A5, A7 | A3: assuming that DRB termination at the gNB is handled/specified, it is not clear why there will be more requirements here than the latency requirement for other UP data over the Uu.  A4: not clear how service continuity is an issue here considering model is terminated at the gNB and data forwarding between gNBs is a fundamental feature of HO. |
| Xiaomi | No need to discuss A2 and A6 (see comments to Q1)  Yes for others | A2: Our understanding is that existing UP security model is applicable for solution 1b therefore security and integrity is supported for solution 1b. |
| TCL | No: A2,  Yes with comments: A6, A3, A7,  Yes: Others | A2: See comments to Q1, Readiness should be“supported”. Given all the solutions support security and integrity, it is useless to compare pros and cons, and suggest removing it.  A6: See comments in Q1.  A3 and A7: See comments in Q1 and Q2-1a; |

**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: |
| Qualcomm | Supports   * Delivery of large models * Delivery of parameter sets or delta models * Lossless model delivery (some clarification may be needed from SA2)   Security and integrity of model | Distributed model storage  May require large storage and processing at gNBs |  |
|  |  |  |  |
|  |  |  |  |

##### 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 2b, support within AMF coverage area based on PDCP status report; For Solution 3b, 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 |  |
| A9 | *(Impact out of RAN2 scope)* |  |
| A10 | No additional gNB impact |  |

**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 |
| Huawei, HiSilicon | No: A3, A7, A5, A6  Yes: Others | **A3, A7**  As we commented for above questions, there should be uniform analysis for them as they are about QoS impacts.  **A5**  For solution 2b/3b, it is FFS whether all of model information are transferred via UP. In the past, some companies commented that some of model information (e.g. model ID, meta data) may be transferred via CP, e.g. from CN to UE, from CN to gNB. In this case, we have the following suggestions:  Readiness: **gNB cannot perform model management directly.**  RAN spec impact: **This column should be only about the impats in Uu interface, so the current wording can be removed as it is not relevant to RAN impacts.**  **A6**  As we commented for above questions, the meaning of A6 is unclear, and what kind of information may need A6 is also unclear. |
| Qualcomm | No for A3, A4, A5, A6  A2, A3, A5, A6, A8 are similar for all solutions, therefore A2, A3, A5, A6, and A8 should be removed. Instead, A9 and A10 should be added. | A3: Do not agree with 2). Not significant as compared to scheduling and Uu delay. Therefore, other latency including forwarding data from CN to gNB should be removed.  A4: Supports retransmission (without any limitations). AF/AS can be connected to multiple AMFs (depending on CN architecture). Need further study from SA2. This is outside RAN2's scope.  A5: gNB can perform model management based on metadata and UE capability signaling. Therefore, even if the model delivery is over UP, the gNB can have full control over model management. Additional signalling may be required over the Uu to indicate the completion of delivery/transfer of the configured model. Remove 2 (in cons) from solution 2b/3b in the pros and cons table. Note that configuration is done based on the UE capability signaling. Model transfer can happen between UE and CN transparent to the gNB; gNB is indicated once model transfer/delivery is complete.  A6: Same comments as in Q1, Q2-1a, Q2-1b. Remove 4 (in cons) from solution 2b/3b in the pros and cons table. |
| Apple | Yes: A1  No: all others | A1 is correct that " No RAN impact".  On A5/A6, we think they are out of RAN2 scope, and thereby RAN2 is not in position in making conclusion (i.e. not capture them in TR). |
| ZTE | Yes for A1, A4, A6 with editorial comments.  No for others | Regarding A4:  For Solution 2b, support within AMF coverage area based on PDCP status report; For Solution 3b, support within LMF coverage area based on LPP signaling segmentation  Regarding A7, please see our comments in above  For other items except for A1, A4, A6, please see our reply in question 1 where all other items seems not be precisely/correctly to reflect the PRO and CONs for each solution. |
| Mediatek | Yes: A1, A3,  No: A2, A4, A5  FFS: A6, A7 A8, A9, A10 | A2  Common for all solutions, which is out of RAN2 scope and can be removed.  A4  Considering the model is delivered through DRB, model transfer/delivery continuity is by default supported, e.g., via AM transmission. But the interaction between gNB and CN/LMF needs to be considered.  A5 Agree with OPPO that we should differentiate 2b and 3b.  A6, A7, A8 see comment in Q1  A9 see comment in Q1  A10 For solution 2a if model management is performed by gNB, there should be some gNB impacts. |
| Interdigital | No (see comment to Q1): A2, A6, A8, A9, A10  No: A4 (see comments),  Yes: A1, A3, A5, A7 | A4: for solution 2b, it is not clear why even more stringent requirement is being put here that we don’t have for normal UP data transmission? |
| Xiaomi | General comment for solution 3b: analysis of solution 3a is applicable to solution 3b.  For solution 2b: No need to discuss A2 and A6 (see comments to Q1), Yes for others | In solution 3b, AI/ML model is transferred between LMF and UE via LPP signalling carried as UP data, as specified in TS 23.273. Therefore in both solution 3a and 3b, AI/ML model are carried in LPP signalling. From RAN2 perspective, there might be no difference between Option 3a and 3b since both options require LPP enhancements to support AI model transfer/delivery.  In summary, our understanding is that analysis of solution 3a is applicable to solution 3b. |
| TCL | No: A2,  Yes with comments: A5, A6, A3, A7,  Yes: Others | A2: See comments in Q1.  A5: It is without RAN2 domain, we agree with Apple‘s comments, and remove the assumption in the “RAN specification impact”.  A6: See comments in Q1.  A3 and A7:See comments to Q1, Q2-1a. |

**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: |
| Qualcomm | Supports   * Delivery of large models * Delivery of parameter set or delta model * Lossless model delivery * Security and integrity of model |  | Some enhancements are required for management.  Indicate gNB when a configure model is downloaded by the UE. |
|  |  |  |  |
|  |  |  |  |

##### 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 |
| A9 | No impact by definition |
| A10 | No additional gNB impact |

**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 |
| Huawei, HiSilicon | No: A3, A7, A6  Yes: Others | **A3, A7**  As we commented for above questions, there should be uniform analysis for them as they are about QoS impacts.  **A6**  As we commented for above questions, the meaning of A6 is unclear, and what kind of information may need A6 is also unclear. |
| Qualcomm | No for A2/A3/A4/A5/A6/A8  A2, A3, A5, A6, A8 are similar for all solutions, therefore A2, A3, A5, A6, and A8 should be removed. Instead, A9, A10 and A11 should be added. | A2: See comments to Q1.  A3: Similar comment as Q2-2b3b. The server can be placed close to gNB to reduce latency.  A4: Same understanding as OPPO.  A5: gNB can perform model management based on metadata and UE capability signaling. Therefore, even if the model delivery is over UP, the gNB can have full control over model management. Additional signalling may be required over the Uu to indicate the completion of delivery/transfer of the configured model. Remove 4 (from cons) from the pros and cons table, as model delivery can happen through DRBs not affecting priority traffic. Note that over-the-top traffic is not provided with high-priority DRBs anyway in the 3GPP network.  A6: See comments to Q1, Q2-1a, Q2-1b.  A8: See comments to Q1. The interoperability issue is similar to any other solution. Remove 2 (from cons) from the pros and cons table, as all models need to satisfy the RAN4 requirements (nothing special to do in this case). |
| Apple | Yes: OK to just capture "No RAN impact" | As we responded in Q1, we do not agree to capture anything on "readiness" in TR, but we are OK to capture " no RAN impact" for solution 4a which is aligned its description " transparent to 3GPP". |
| ZTE | Yes for A1,A4, A6  No for all others | For all other items except for A1,A4,A6, please see our reply in question 1 where all other items seems not be precisely/correctly to reflect the PRO and CONs for each solution. |
| Mediatek | Yes: A1, A3,  No: A2, A4, A5  FFS: A6, A7,A8, A9, A10  We should soften the statement ‘No RAN impact’. It may have RAN impact depends on to what extend RAN awareness is required for model delivery, especially when the AI/ML model doesn’t have good generalization performance, e.g., cell-specific model. | A2  Common for all solutions, which is out of RAN2 scope and can be removed.  A4  Considering the model is delivered through DRB, model transfer/delivery continuity is by default supported, e.g., via AM transmission.  A5  If network controllability is for model management, it is the same for all solutions, e.g., NW needs to perform model monitoring and model control.  If network controllability is for model delivery, it’s for sure that NW doesn’t have direct control on model delivery. But considering the case that the AI/ML model may not have good generalization performance, which is associated to certain RAN configuration, condition, scenarios, site, etc, certain RAN involvement may be needed.  A6, A7, A8 see comment in Q1  A9, A10  If the model doesn’t have good generalization performance, e.g., area-specific model, RAN involvement will be helpful instead of downloading the model blindly. |
| Interdigital | No (see comment to Q1): A2, A6, A8, A9, A10  No: A4 (see comments)  Yes: A1, A3, A5, A7 | A4: not clear why UP service continuity cannot support this. |
| Xiaomi | No need to discuss A2 and A6 (see comments to Q1)  No for A4  Yes for others | **A4**: in our understanding, the model transfer continuity across gNBs is just like handling of normal data during handover, and it is supported by default. |
| TCL | No: A2,  Yes with comments: A6, A3, A7  Yes: Others | A2: See comments to Q1;  A6: See comments to Q1;  A3 and A7, It is unclear to us about the model transfer/delivery between the OTT server and UE, anyway, the upper-layer data may also influence the DRB and QoS priority setting, same comments with 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: |
| Qualcomm | Supports   * Delivery of large models * Delivery of parameter set or delta model * Lossless model delivery * Security and integrity of model |  | Some enhancements are required for management.  Indicate gNB when a configure model is downloaded by the UE. |
|  |  |  |  |
|  |  |  |  |

##### 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 |  |
| A9 | *(Impact out of RAN2 scope)* |  |
| A10 | No additional gNB impact |  |

**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 |
| Huawei, HiSilicon | No: A1, A3, A7, A6  Yes: Others | **A1**  For solution 4b, the tranmission path is unclear to us. It may include:  UP: OAM (as a server) to UE  CP: OAM to entity X, and then to UE. Entity X may be: CN, gNB, others?  **To us, RAN2 just agreed to split Solution 4, and there are lots of aspects to be clarified first for Solution 4b.**  For now, the readiness text seem to be based on some assumptions on the tranmission path, and we suggest to make it clear. For example:  Readiness: **over CP, if the transmission path is “OAM -> RAN -> UE” and CP is used for RAN -> UE, not supported. Over UP: supported.**  RAN spec impact: **over CP, if the transmission path is “OAM -> RAN -> UE” and CP is used for RAN -> UE, if model size larger than 45kBytes, extend RRC segment number**  **A3, A7**  As we commented for above questions, there should be uniform analysis for them as they are about QoS impacts.  **A6**  As we commented for above questions, the meaning of A6 is unclear, and what kind of information may need A6 is also unclear. |
| Qualcomm | No for A1, A2, A3, A5, A6, A8  A2, A3, A5, A6, A8 are similar for all solutions, therefore A2, A3, A5, A6, and A8 should be removed. Instead, A9, A10 and A11 should be added. | A1: Same view as OPPO  A2, A3, A5, A6, A8: Similar comments as previous. |
| Apple | Yes: A2 (i.e. Note: The details security and integrity of solution 4b is out of RAN scope)  No: all others | As we responded in Q1, we do not agree to capture anything on "readiness" in TR, but we are OK to capture " (i.e. Note: The details security and integrity of solution 4b is out of RAN scope)" for solution 4b.  On A1: same view as OPPO. We don't prefer to capture it in TR.  On A5/A6, we think they are out of RAN2 scope, and thereby RAN2 is not in position in making conclusion (i.e. not capture them in TR). |
| ZTE | Comments for A1  Yes for A4, A6  No for all others | Regarding A1, we think the data transmission between UE and OAM has not been supported yet, regardless of UP and CP.  For A1:  Readiness: CP based solution: not support, UP based solution: not support  Specification impact: For UP based solution, NW shall at least provide IP address of OAM to UE. For CP based solution，extension of RRC segmentation may be needed.  Regarding A7 see our comments in above  For other items except for A1, A4, A6, please see our reply in question 1 where all other items seems not be precisely/correctly to reflect the PRO and CONs for each solution. |
| Mediatek | Many aspects are out of RAN scope: A1, A2, A3, A4, A5, A7, A9, A10, | A1  It is made based on the assumption that there is a transmission path between UE and OAM via gNB and both CP/UP based transmission are supported. But we are not sure whether the assumption is correct or not.  If I understand correctly, the solution 4 is a UP based solution and the transmission tunnel may be transparent to 3GPP. **Solution 4: Server (e.g. OAM, OTT) can transfer/delivery AI/ML model(s) to UE (e.g. transparent to 3GPP).**  Not sure whether CP based transmission should be considered. From network implementation point of view, if the model is transmitted from OAM to UE via CP, there is not much difference from solution 1a?  A2  Common for all solutions and need to be removed  A3  Agree with OPPO  A4, A5, A7, A9, A10 out of RAN2 scope.  FFS on A6, A8 |
| Interdigital | No (see comment to Q1): A2, A6, A8, A9, A10  No: A1 (see comments)  Yes: A3, A4, A5, A7 | A1: not clear why the CP based solution doesn’t support bigger sizes (if segmentation is done at the OAM) |
| Xiaomi | No need to discuss A2 and A6 (see comments to Q1)  Yes for others |  |
| TCL | No: A1, A2,  Yes with comments: A6, A3, A7,  Yes: Others. | A1: We agree the OPPO ‘s comments.  A6, A3, A7: Similar comments as previous. |

**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)