**3GPP TSG-RAN RAN2 #121 R2-23xxxxx**

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

**Agenda Item:**  **8.16.2 AIML methods**

**Source: Huawei (email rapporteur)**

**Title:** **Summary of 027 pros and cons of the listed solutions (Huawei)**

**Document for: Discussion and Decision**

# 1 Introduction

This is the email report of [AT121][027].

* Offline 027 (Huawei) attempt a first round of capturing expected pros and cons of the listed solutions.

Deadline: 22:00 UTC, Thursday March 2nd

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

|  |  |  |
| --- | --- | --- |
| **Company** | **Name** | **Email Address** |
| Qualcomm | Rajeev Kumar | rkum@qti.qualcomm.com |
| Apple | Peng Cheng | pcheng24@apple.com |
| Huawei, HiSilicon | Jun Chen | jun.chen@huawei.com |
| OPPO | Jiangsheng Fan | fanjiangsheng@oppo.com |
| Interdigital | Oumer Teyeb | [Oumer.teyeb@interdigital.com](mailto:Oumer.teyeb@interdigital.com) |
| Intel | Ziyi Li | Ziyi.li@intel.com |
| LGE | Soo Kim | soo.kim@lge.com |
| NTT DOCOMO | Souki Watanabe | souki.watanabe.gf@nttdocom.com |
| NEC | Xuelong Wang | xuelong.wang@emea.nec.com |
| Nokia | Jerediah Fevold | jerediah.fevold@nokia.com |
| CMCC | Jiayao Tan | tanjiayao@chinamobile.com |
| Samsung | Chadi Khirallah | [c.khirallah@samsung.com](mailto:c.khirallah@samsung.com) |
| vivo | Boubacar Boubacar | kimba@vivo.com |
| Lenovo | Congchi Zhang | zhangcc16@lenovo.com |
| Spreadtrum | Xiaoyu Chen | xiaoyu.chen@unisoc.com |
| ZTE | Fei Dong | Dong.fei@zte.com.cn |
| Ericsson | Felipe Arraño Scharager | felipe.arrano.scharager@ericsson.com |
| CATT | Erlin Zeng | erlin.zeng@CATT.CN |

# 2 Discussion

In the report [2], the pros/cons of the listed solutions have been summarized in relevent sections, so it is suggested to directly use them for this email discussion.

According to the chair notes, the following solutions are listed:

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).

Since the email scope is to attempt a first round of capturing pros/cons, the email rapporteur has the following suggestions:

* For each bullet for pros/cons, it will be good to see companies’ preferences, e.g. yes/no. More comments are also welcome
* Based on the companies’ views, there will be a decision for each bullet, i.e. agreeable, FFS
* In the end of this email discussion, the expected outcome is that we will have a full list, and some bullets for pros/cons are agreeable, and some bullets are FFS (based on the majority views). The next step may be that RAN2 can address FFS and have more technical discussions on how to solve them

## 2.1 Discussion on pros/cons of Solution 1a

The pros/cons of Solution 1a are listed in the table below and they are exactly the same as pros/cons listed in [2]. The numbering is put.

|  |  |
| --- | --- |
| **Solution 1a** | gNB can transfer/deliver AI/ML model(s) to UE via RRC signalling |
| **Pros** | 1. No inter-operability issues 2. The gNB can transfer/delivery the models to UE with limited latency. ~~Can be less latency compared with other solutions. Some companies think Solution 1a can be flexible, as different SRBs can meet different transmission requirements~~ 3. Can be higher priority compared with model transfer/delivery via UP/DRB 4. ~~If the model is visible to the RRC layer, delta configuration can be used to reduce the signalling overhead~~ (merged to 6) 5. Allows existing UE context transfer from source to target to be applicable for mobility 6. The existing RRC signaling solutions can be reused as baseline, at least including delta signaling and segementation 7. SRB transmission is generally more robust than DRB (assuming gNB is not aware of AI/ML model transfer in one DRB as in legacy) 8. Limited specification impact for supporting transfer/delivery of a model with a few KB in size 9. Additional security and verification may not be necessary as the UE already established security before the transfer is initiated 10. Attached metadata to the transfer/delivery process is synchronized with the transfer/delivery process 11. gNB can take the control of the AIML model transfer itself, which can not be achieved by traditional UP based solution |
| **Cons** | 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 4. Some companies think that it worths to clarify whether the model is generated by NG-RAN or not. If the model is generated by upper layer and transmit to NG-RAN within network, some of the drawbacks listed by companies above does not exist, e.g. service continuity, etc 5. gNB would have to store all the models for delivery 6. May require massive update of existing gNBs to support ML functionalities 7. ~~For overhead, at RRC layer, if there are some RRC segments, it may introduce some overhead. For the overhead below RRC, there are not much differences between CP-based and UP-based solutions~~ (merged to Bullet 1/2)   New bullet:  If *RRCReconfiguration* message is reused to deliver AI model, the transmission of legacy radio configuration may be delayed due to large AI model delivery. |

The email rapporteur suggests to collect companies’ views/preferences on the above bullets, e.g.:

* For Pros, Yes to all bullets or bullet 1, 2, 3, …; No to all bullets or bullet 1, 2, 3, …
* For Cons, Yes to all bullets or bullet 1, 2, 3, …; No to all bullets or bullet 1, 2, 3, …
* Companies can also provide more comments, e.g. merge some bullets, add more bullets, improve the existing wording, provide more technical comments

**Q1: What are the comments on pros/cons of Solution 1a?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Comments on Pros** | **Comments on Cons** |
| Qualcomm | No –   1. we do not think the interoperability issue is associated with whether CP or UP-based method is used. It has more to do with who trains the model. If the model is developed by the UE, the gNB needs to know where the model is applicable irrespective of whether CP/UP is used. 2. we do not agree with “Can be less latency compared with other solutions.” Other solutions like 1b can achieve similar latency. Also, for 1a, there is additional RRC processing latency. 3. We have DRB supporting delay-critical applications. So not sure if this is true. 4. Model visibility at the RRC is not the requirement for supporting delta configuration. 5. Existing UE transfer is used for coordination of configuration between gNB. SRB is reset during the mobility. This is the new use case, therefore not sure if this statement is true   9- Additional security and verification is required for ensuring that model is made available to the right UEs.  10- What is meta data, is the meta info? If yes, then meta info is for control and management of the model, not required for model delivery.  May be –  7 (only valid with the assumption)  Yes – 8, 11 | Yes – 1 to 7. |
| Apple | Yes for all as starting point of next round of discussion  - We don't prefer to repeat the same discussion again in previous post-meeting email discussion.  - Our understanding is that the list is just used for starting point of next round discussion, to facilitate discussion.  - In our understanding, the list is drafted based on majority view, which should be respected. We do not see any necessity to preclude any point at this stage. | Yes for all as starting point of next round of discussion  - We don't prefer to repeat the same discussion again in previous post-meeting email discussion.  - Our understanding is that the list is just used for starting point of next round discussion, to facilitate discussion.  - In our understanding, the list is drafted based on majority view, which should be respected. We do not see any necessity to preclude any point at this stage. |
| Huawei, HiSilicon | 10: Suggest to put FFS for it as metadata and how to transfer it need more discussions.  Others are ok. | Ok for all. |
| Xiaomi | Comments on,  2- 1a and 1b should expect similar delay  3- not sure this is cons. AI model does not necessarily require higher priority than all data.  11- I guess we can agree traditional UP based solution is not feasible. This bullet can be removed. | Additional bullet,  If *RRCReconfiguration* message is reused to deliver AI model, the transmission of legacy radio configuration may be delayed due to large AI model delivery. |
| OPPO | No to bullet 1   * For bullet 1, based on RAN1 latest agreement during this meeting, both open format and proprietary format are considered for model transfer/delivery, if proprietary format is used for solution1a, a RRC container generated by gNB, e.g. RRC layer, may be used to do encapsulation for AI model, in this case, inter-operability issue may still be there;   Rewording to bullet 2   * It’s sufficient to just keep the first sentence; * The second sentence may be incorrect compared to solution1b as data is transmitted continually via DRB, which can be comparable with solution1a; * The third sentence is also not needed, as we also have AM/UM mode and/or QoS control for DRB solution, hard to say only SRBs can meet different transmission requirements;   Comments to bullet 4   * Can be merged with bullet 6, it says the same thing;   No to bullet 5   * This is wrong, source cell will not transfer the non-confirmed RRC message to the target based on current spec;   No to bullet 7   * As mentioned in bullet1, if proprietary format is used for solution1a, a RRC container generated by gNB, e.g. RRC layer, may be used to do encapsulation for AI model, in this case, gNB is also not aware of the AI model, we really don’t understand what ‘robust’ refers to.   No to bullet 8   * In current stage, bullet 6 is sufficient, consider the potential enhancement for mobility, the impact may not be that limited;   No to bullet 9   * All the solutions have security during transmission, why this is a pros for 1a?   No to bullet 11   * Bullet 10 is sufficient, if added, solution1b should be added also. |  |
| Interdigital | OK for all | Ok for all |
| Intel | Yes to all | Yes - 4.  No - 1. 2. 5. 6. 7.  In our understanding, in this solution, it is also possible that the model is first generated by upper layer, e.g. OAM, etc. Therefore, the segmentation can happen at upper layer, rather than at RRC. Hence, RRC message size limitation may not be the limitation.  Similar as bullet 4, we can add it as an exceptional case for 1 and 2.  Another suggestion is to summarize all limitation caused by RRC segmentation into a single bullet, as a consequence of RRC message size limitation, e.g. merge bullet 2/5/6/7 into bullet 1.  No – 3.  For the support of mobility, we are wondering whether it is really needed. The AI/ML model normally suits for a certain environment, based on the input data it uses for model training. It is possible that one model performs good at source gNB, but performance of the same model drops at target gNB after mobility. Hence, we think it worth to discuss the necessarily of supporting handover/mobility and not capture mobility related cons at this stage. |
| LGE | It seems that the case where the model is visible (e.g., RRC IE) in RRC and the case where it is not visible (e.g. encapsulated msg) in RRC are mixed. We think pros will be different depending on what format it is. | Yes to all bullets |
| NTT DOCOMO | Yes to all bullets. | 1. About wording   Face challenges to convey large “or no upper limit” size AI model by RRC message (e.g. >45kBytes) |
| NEC | Yes – 1 to 11 | Yes – 1 to 2.  For 3, we think there is a possibility to resume the model transmission 8 for remaining part of the model, instead of restarting the whole model transfer. |
| Mediatek | General comments:   1. Certain pros/cons come from CP/UP comparison, while other pros/cons are due to from where the model is delivered/transferred. It would be good to mention that compared with which/what kind of solutions, the solution concerned has pros/cons. 2. We should consider the metrics that really matters instead of listing all of them.   Yes:  2. The gNB can transfer/delivery the models to UE with less latency than solution 2/3/4.  4. Support delta configuration and reduce signaling overhead  5. The benefit of ‘Allows existing UE context transfer from source to target to be applicable for mobility’ is to reuse current mobility mechanisms with less specification change.  9, 11  FFS:   1. Agree with QC that the inter-operability issue has more to do with where the model is trained. It’s also related to in what kind of format the model is transferred/delivered. 2. Not sure whether having higher priority than DRB is a pro that really matters. 3. This may be redundant with other aspects, e.g.,4 and 5. 4. It may not be true under certain assumption. 5. UP solution may have the least specification impact. 6. Attached metadata to the transfer/delivery process is synchronized with the transfer/delivery process | Yes  1, 2, 5, 6  FFS:  3. Depends on how large the model size is.  4. 7. |
| Nokia, Nokia Shanghai Bell | **Agree**  10. True, but it is FFS what, if any, metadata is part of the model transfer/delivery.  11. True  **Disagree**  1. It isn’t clear what is meant by interoperability issues. As stated by QC, one interpretation could refer to compatibility of a model with a device.  2. DRBs can be configured with low latency as well.  3. DRBs can also be set with high priority.  4. We do not think that this table is the right place to discuss delta configurations as we haven’t defined what that means.  5. It isn’t clear what context related to the model transfer/delivery is.  6. It should be clarified what delta signaling means in the context of model transfer/delivery.  7. DRBs also transmit data in a robust manner, with automatic adaptation and retransmissions. What is unclear is what is gained by the perceived additional robustness of the CP for what is essentially a file transfer.  8. The specification impact needs to be studied regardless of whether the model fits inside of the RRC message.  9. Clarification is needed to what type of security (RRC?), and what verification means. | **Agree** 1 – 3  **Disagree**  4. From the UE’s perspective, the transfer/delivery is from the gNodeB, so any transfers from NFs to the gNodeB don’t need to be considered here.  5. We do not see the need to set requirements on where models are stored.  6. gNodeB updates are expected for any new features. This isn’t a con specific to AIML transfer/delivery. |
| CMCC | OK for all as starting point | OK for all as starting point |
| Samsung | Ok for all as a starting point for the next round of discussion. | Ok for all as a starting point for the next round of discussion. |
| vivo | Ok for all as starting point | Bullet 4 may require some clarification  From our understanding this “Bullet 6” may be valid to all other solution, i.e., this is not a particular drawback to only this solution  Bullet 7: this bullet seems to overlap with Bullet 1/2 |
| Lenovo | They look in general ok to us.  For those seem controversial, we feel they depend on the deployment scenarios. Prefer to take them as start point, and further discuss when scenario becomes more clear. | They look in general ok to us.  For those seem controversial, we feel they depend on the deployment scenarios. Prefer to take them as start point, and further discuss when scenario becomes more clear. |
| Spreadtrum | Comments on bullet 1:  Not sure. For the current use case in SI, RRC needs to deliver the received AIML model to PHY layer. Some inter-operability may still be needed.  Comments on bullet 3:  In case it is time-sensitive model transfer/delivery and with small model size. | Yes to all |
| ZTE | * Regarding the item 3:   + “Can be higher priority compared with model transfer/delivery via UP/DRB” It shall be removed from the PROs, the priority of the Radio bearer is mainly relying on the LCH priority, the DRB can be configured with a higher priority than an SRB. * Regarding the item 6:   + Is the delta signalling the same thing with the item 4, if it is , this redundant with item 4. And, we do not think the segmentation mentioned in item 6 is a PRO compare to any UP solution since UP data can be segmented, so item 6 can be removed. * Regarding the item 7:   + We are a little bit confusing about the wording ‘robust’? Does it mean reliability? If it is, both DRB and SRB can support PDCP duplication (DL).   + It is also not clear why we need an assumption of gNB is not aware of the model transfer.   Yes to others | * Regarding the item5:   + AI model transfer always need the terminated point to store the AI model, it is not clear the reason why it is considered as a CONs…..   Yes to others. |
| Ericsson | No:   * 2. Not clear to us the “other solutions” so prefer not to capture it. * 6. Well there would be a need to discuss on segmentation for this purpose.   OK to take the others as baseline. We should always be able to come back to the table and update/modify it. | No:  5. We don’t see a need to assume this.   OK to take the others as baseline. We should always be able to come back to the table and update/modify it. |
| CATT | Yes to bullet 1-7, 9, 11.  For bullet 8, to be compared with solution 1b, there has limited specification impact, but to be compared with solution 2a/2b, solution 1a has same or larger specification impact.  For bullet 10, we should firstly decide what the metadata is and how to transfer it. | Yes to bullet 1, 2, 3, 6, 7.  For bullet 4, indeed we should clarify whether the model is generated by NG-RAN or not firstly, but it is not Cons.  For bullet 5, if the model is generated by NG-RAN, it is needed for gNB to store them. It is not a Cons. |

## 2.2 Discussion on pros/cons of Solution 2a and 3a

The pros/cons of Solution 2a are listed in the table below and they are exactly the same as listed in [2]. The numbering is put.

In the report [2], it is observed that the analysis for Solution 2a can be also used for Solution 3a. So it is proposed to discuss pros/cons for both solutions.

|  |  |
| --- | --- |
| **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 |
| **Pros** | 1. No inter-operability issues 2. If the model is visible to the NAS layer, delta configuration can be used to reduce the signaling overhead 3. SRB transmission is generally more robust than DRB (assuming gNB is not aware of AI/ML model transfer in one DRB as in legacy) 4. Model management like model update and model sharing procedure is easy compared with Solution 1a 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) 7. ~~Some companies think~~ Solution 2a can be flexible, as different SRBs can meet different transmission requirements 8. ~~Dela configuration may be possible, when only model parameter update is required~~ (merged to 2)   New:  The control plane makes it possible to transmit information along with the model.  For bullet 2, it shall be NAS/LPP layer |
| **Cons** | 1. Face challenges to convey large size or “no upper limit size” AI model by RRC message (e.g. >45kBytes) 2. Larger latency compared with Solution 1a 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 |

**Q2: What are the comments on pros/cons of Solution 2a and Solution 3a?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Comments on Pros** | **Comments on Cons** |
| Qualcomm | Yes -  4, 5, 6, 7, and 8.  No -   1. We do not think the interoperability issue is associated with whether CP or UP-based method is used. It has more to do with who trains the model. If the model is developed by the UE, the gNB needs to know where the model is applicable irrespective of whether CP/UP is used. 2. For delta configuration, model visibility is not the requirement. 3. No. | Yes –  Agree with 1 to 4. |
| Apple | Yes for all as starting point of next round of discussion  - We don't prefer to repeat the same discussion again in previous post-meeting email discussion.  - Our understanding is that the list is just used for starting point of next round discussion, to facilitate discussion.  - In our understanding, the list is drafted based on majority view, which should be respected. We do not see any necessity to preclude any point at this stage. | Yes for all as starting point of next round of discussion  - We don't prefer to repeat the same discussion again in previous post-meeting email discussion.  - Our understanding is that the list is just used for starting point of next round discussion, to facilitate discussion.  - In our understanding, the list is drafted based on majority view, which should be respected. We do not see any necessity to preclude any point at this stage. |
| Huawei, HiSilicon | Ok for all. | Ok for all. |
| Xiaomi |  | Regarding solution 3a, No for 4. For AI/ML based positioning, LMF is feasible for performing AI/ML model LCM since currently the LMF is responsible for managing positioning including positioning method determination, PRS resource configuration, aperiodic and Semi-persistent SRS activation/deactivation and so on. |
| OPPO | Maybe No to bullet 1   * The similar reason as Q1 bullet1;   No to bullet 3   * The similar reason as Q1 bullet7   we really don’t understand what ‘robust’ refers to.  Maybe No to bullet 7   * Can someone clarify why this is the advantage of SRB, but not DRB, it seems unclear from our side;   Comments to bullet 6   * I don’t know whether It’s also applied for solution3a as LPP spec is in RAN2 domain;   No to bullet 8  Can merge with bullet 2; |  |
| Interdigital | Ok for all, but maybe 2 & 8 can be combined into one (e.g., delta update is possible) | Ok for all |
| Intel | Yes – 1,2,3,4,5,6,8  No – 7  NAS is using SRB2 which is a lower priority SRB, it’s not clear to us how to understand “different SRBs” to meet different transmission requirement by using SRB2. Does it mean, in this solution, another SRB is used to carry NAS signaling? | Yes to 2.3.4  No – 1 see our comment above on RRC segmentation |
| LGE | Yes to all bullets | Yes to all bullets |
| NTT DOCOMO | Yes to all bullets. | 1. About wording   Face challenges to convey large “or no upper limit” size AI model by RRC message (e.g. >45kBytes) |
| NEC | Yes for 1,3, 4, and 5.  For 2, it is not clear what is delta configuration? | Agree with 1 to 4.  In addition, the solution actually make a restriction to ask all of the UEs managed by one 5GC to apply a single AIML model. |
| Mediatek | Yes: 2, 4, 5  FFS: 1, 3, 7   1. Not sure whether RAN2 impact can be considered. Generally, we should consider the overall impact in both RAN2 and other WGs e.g., SA.   Item 8 is the same as 2? | Yes for all. |
| Nokia, Nokia Shanghai Bell | **Note:** We should respect the structure we came up with as a group and treat 2a and 3a separately.    **Disagree**   1. It isn’t clear what is meant by interoperability issues. As stated by QC, one interpretation could refer to compatibility of a model with a device. 2. We do not think that this table is the right place to discuss delta configurations as we haven’t defined what that means. 3. DRBs can also be set with high priority. 4. It isn’t clear why model management is easier from NAS or LPP than from RRC, and we shouldn’t discuss model update and model sharing here.   6. Agree for NAS, but for LPP signalling, this is in RAN2 scope.  7. The limitation on SRB use for cases 1a, 2a, and 3a should be clarified.  8. We do not think that this table is the right place to discuss delta configurations as we haven’t defined what that means.  **Additional Pro**  We think that the control plane makes it possible to transmit information along with the model. | **Agree –**  1. True for large models  3. True in addition to potential RRC segmentation.  **Disagree** –  2. Low latency is possible with UP solutions.  4. While this might be true, actions other than transfer/delivery are not in scope of this table. |
| CMCC | OK for all as starting point | Comment on 4, share similar view with Xiaomi.  We suggest to change CN to CN (expect LMF). |
| Samsung | Ok for all as a starting point for the next round of discussion. | Ok for all as a starting point for the next round of discussion. |
| vivo | Although 2a and 3a share most pros/cons, solution 3a can benefit from the existing framework of positioning. Therefore, we propose to add another pro:  For bullet 2, it shall be NAS/LPP layer | Ok for all as starting point |
| Lenovo | OK for all | OK for all |
| Spreadtrum | Comments on bullet 1:  Not sure. For the current use case in SI, RRC needs to deliver the received AIML model to PHY layer. Some inter-operability may still be needed. | Yes to all |
| ZTE | * Regarding item 3:   + The same comments as the response in Q1;   + We are a little bit confusing about the wording ‘robust’? Does it mean reliability? If it is, both DRB and SRB can support PDCP duplication (DL).   + It is also not clear why we need an assumption of gNB is not aware of the model transfer.   The item 2 and item 8 seems redundant, can remove one of them.  OK to others. | OK for all |
| Ericsson | No:  - 2. Not clear how this has been deducted. - 4. Not clear how this has been deducted. - 8. Not clear how this has been deducted.  OK to take the rest as starting point. | OK to take cons provided as starting point. |
| CATT | We do not think the solution 2a and 3a can be discussed in the same table. Since solution 2a can be mapped to use cases of CSI feedback enhancement and BM (model inference performed in UE and gNB), but solution 3a can be mapped to Positioning accuracy enhancement (model inference performed in UE and LMF). We need to discuss the Pros and Cons for the specific use cases.  Yes to bullet 1-3, 6, 8.  For bullet 4, we think the Model management is similar with Solution 1a, they both use control signalling for e.g. model update.  For bullet 5, we think we should clarify whether the model is stored in NG-RAN or in CN.  For bullet 7, almost only SRB2 can be used if SRB2 is established. | Yes to bullet 2, 4.  For bullet 1, there can have segmentation in NAS layer.  For bullet 3, no matter to use NAS or AS solution, the overhead is always introduced. |

## 2.3 Discussion on pros/cons of Solution 1b

The pros/cons of Solution 1b are listed in the table below and they are exactly the same as listed in [2]. The numbering is put.

|  |  |
| --- | --- |
| **Solution 1b** | gNB can transfer/deliver AI/ML model(s) to UE via UP data |
| **Pros** | 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 3. ~~Can handle model delivery/transfer during mobility efficiently~~ (removed as it contradicts with 5) 4. Suitable for transferring multiple models simultaneously 5. Compared with CP-based solutions, it may not need to consider CP message segmentation, CP message blocking issue |
| **Cons** | 1. [FFS] more specification impacts~~Impacts due to new solutions (need more discussions as the solution details are not clear for now)~~ 2. gNB will control the AI model transfer/delivery session setup/release, which is usually controlled by CN in traditional procedure 3. For the delay analysis, it depends on the principle and basic flow of Solution 1b 4. RRC layer may not comprehend the model content, and the gNB may not perform delta-model transfer/delivery based on current user plane framework 5. Not compatible with current mobility procedure. Supporting model transfer during mobility is not so straightforward 6. DRB transmission is generally less robust than SRB (assuming gNB is not aware of AI/ML model transfer in one DRB as in legacy)   New:  if the model is stored in the CU, the F1 interface overhead is inevitable. |

**Q3: What are the comments on pros/cons of Solution 1b?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Comments on Pros** | **Comments on Cons** |
| Qualcomm | Yes – 1 to 5. | Yes –  No –  4, 5, and 6.  Not sure –  What 2 means?  Not sure if 1, 3 is cons. |
| Apple | Yes for all as starting point of next round of discussion  - We don't prefer to repeat the same discussion again in previous post-meeting email discussion.  - Our understanding is that the list is just used for starting point of next round discussion, to facilitate discussion.  - In our understanding, the list is drafted based on majority view, which should be respected. We do not see any necessity to preclude any point at this stage. | Yes for all as starting point of next round of discussion  - We don't prefer to repeat the same discussion again in previous post-meeting email discussion.  - Our understanding is that the list is just used for starting point of next round discussion, to facilitate discussion.  - In our understanding, the list is drafted based on majority view, which should be respected. We do not see any necessity to preclude any point at this stage. |
| Huawei, HiSilicon | Ok for all. | 1: suggest to keep FFS, as how Solution 1b works is unclear in RAN2, and then there may be more and more disadvantages later (if more details will be figured out)  2: We understand that this bullet means extra work for interfaces between CN and gNB, which is a disadvantage from our point of view.  Others are ok. |
| Xiaomi |  | Comments on,  2- There is no AI model transfer/delivery session. |
| OPPO |  | Comments to bullet 6   * The similar reason as Q1 bullet7   we really don’t understand what ‘robust’ refers to. |
| Interdigital | OK to all | Some of the cons need rephrasing, e.g.  #1: better to say “more specification impact”?  #2: (is that really a con?)  #3: it is not clear. We may need to rephrase “likely more latency than CP based solution”  #4: can we rephrase “delta update may not be feasible”?  #5:can we clarify a bit why it is not compatible with current mobility? Maybe from network pov, but not really from UE’s pov, right? |
| Intel | Yes – 1.2.4  No – 3 see our comment to handover/mobility support in solution 1a  No – 5 see our comment to RRC segmentation in solution 1a | Yes to 1.2.3.4.6  No – 5 see our comment to handover/mobility support in solution 1a |
| LGE | Yes to all bullets | Not sure if 2 and 3 are cons.  Others are ok. |
| NTT DOCOMO | Yes to all bullets. | Yes to all bullets. |
| NEC | Yes for 1 and 5.  For 2, the overhead over air interface is the same comparing with CP based solution. | No for 1-6. |
| Mediatek | Yes: 1, 2, 4, 5  FFS: 3, it depends on where the model is terminated and the solutions. | Yes: 2, 4, 5  FFS: 1,3, 6 |
| Nokia, Nokia Shanghai Bell | **Agree (but wording should be revised)** 1, 2, 4, 5  Pro #3 contradicts con #5 | **Agree**  5. How to resolve mobility issues are FFS. It isn’t clear what will be used to initiate gNodeB UP data transfer, and what role the CP would play in a UP data transfer.  **Disagree**  1. We think this applies to all new procedures, generally.  2. It needs to be clarified what control plane messages are needed to facilitate the UP data transfer from the gNodeB.  3. Analysis is expected, but it isn’t a con related to UP data transfer.  4. We do not think that this table is the right place to discuss delta configurations as we haven’t defined what that means.  6. DRBs also transmit data in a robust manner, with automatic adaptation and retransmissions. What is unclear is what is gained by the perceived additional robustness of the CP for what is essentially a file transfer. |
| CMCC | OK for all as starting point | OK for all as starting point |
| Samsung | Ok for all as a starting point for the next round of discussion. | Ok for all as a starting point for the next round of discussion. |
| vivo | Bullet 3 in the pros seems to contradict bullet 5 in the cons. | Ok for all as starting point |
| Lenovo | OK for all | OK for all |
| Spreadtrum | Yes to all for a good start  Comments to Bullet 1: The assumption on configuring different 5QI for AIML model transfer rely on the mechanism how solution 1b works. | Yes to all |
| ZTE | Regarding the item 2:  Why Control plane overhead is more serious than the UP plane overhead? To our understanding, the signaling overhead is always there no matter which plain is used. | A new item can be added  if the model is stored in the CU, the F1 interface overhead is inevitable. |
| Ericsson | OK as baseline | Is 2 a Con? |
| CATT | Yes to bullet 1, 2, 5.  For bullet 3, to use UP solution, we do not think it can be efficient to handle model delivery/transfer during mobility.  For bullet 4, why we need to transfer multiple models simultaneously? | Yes to all bullets. |

## 2.4 Discussion on pros/cons of Solution 2b and 3b

The pros/cons of Solution 2b are listed in the table below and they are exactly the same as listed in [2]. The numbering is put.

In the report [2], it is observed that Pros of Solution 2b is the same as the Pros of Solution 1b. And it is observed that the analysis for Solution 2b can be also used for Solution 3b. So it is proposed to discuss pros/cons for both solutions together.

|  |  |
| --- | --- |
| **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 |
| **Pros** | 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, Solution 2b and 3b~~this Solution 1b~~ can reduces control plane overhead, reduces overhead at gNB for model delivery/transfer 3. ~~Can handle model delivery/transfer during mobility efficiently~~ (removed as it contradicts with 5) 4. Suitable for transferring multiple models simultaneously 5. Compared with CP-based solutions, it may not need to consider CP message segmentation, CP message blocking issue |
| **Cons** | 1. It may have inter-operability issues 2. CP signalling is needed to configure and initiate the model transfer from the CN 3. The AI model transfer/deliver has more delay and is less robust compared with Solution 1a 4. May be unable to support delta-model transfer/delivery based on current user plane framework 5. Not compatible with current mobility procedure. Supporting model transfer during mobility is not so straightforward 6. DRB transmission is generally less robust than SRB (assuming gNB is not aware of AI/ML model transfer in one DRB as in legacy) |

**Q4: What are the comments on pros/cons of Solution 2b?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Comments on Pros** | **Comments on Cons** |
| Qualcomm | Yes –  1 through 5. | Yes –  No –  1 - We do not think the interoperability issue is associated with whether CP or UP-based method is used. It has more to do with who trains the model. If the model is developed by the UE, the gNB needs to know where the model is applicable irrespective of whether CP/UP is used.  2 – RRC can configure UE to download model if not available at the UE. Meta info at the gNB can be used for this purpose  3 – we do not think it is less robust. The 5GC knows about the model, it can provide desired QoS requirement. Latency wise, we think we need to consider RRC processing delay. Therefore, we believe that it may not higher overall delay.  4 – delta model transfer can be supported. The gNB can use meta info for doing this.  5 – No.  6 - we do not think it is less robust. The 5GC knows about the model, it can provide desired QoS requirement |
| Apple | Yes for all as starting point of next round of discussion  - We don't prefer to repeat the same discussion again in previous post-meeting email discussion.  - Our understanding is that the list is just used for starting point of next round discussion, to facilitate discussion.  - In our understanding, the list is drafted based on majority view, which should be respected. We do not see any necessity to preclude any point at this stage. | Yes for all as starting point of next round of discussion  - We don't prefer to repeat the same discussion again in previous post-meeting email discussion.  - Our understanding is that the list is just used for starting point of next round discussion, to facilitate discussion.  - In our understanding, the list is drafted based on majority view, which should be respected. We do not see any necessity to preclude any point at this stage. |
| Huawei, HiSilicon | Ok for all. | Ok for all. |
| OPPO |  | No to bullet 1   * We think inter-operability issues may be common for all 3GPP visible model transfer solution no matter CP or UP solution. |
| Interdigital | OK for all | OK for all except 5 (similar comment to 5 as in Q3) |
| Intel | Yes to all | Yes to all |
| LGE | Yes to all bullets | Yes to all bullets |
| NTT DOCOMO | Yes to all bullets. | Yes to all bullets. |
| NEC | Yes for 1 - 5. | Yes for 2,3,4 and 6. |
| Mediatek | Yes: 1, 2, 4, 5  FFS: 3, how to make the model visible to gNB needs to be considered if LCM is controlled by gNB. | Yes: 2, 4, 5  FFS: 1, 3, 6 |
| Nokia, Nokia Shanghai Bell | **Note:** We should respect the structure we came up with as a group and treat 2b and 3b separately.  Agree, but the wording should be revised for points 1, 4, 5  2. This should show solution 2b/3b.  Pro #3 is contradictory with Con #5 | **Agree**  5. How to resolve mobility issues are FFS. It isn’t clear what will be used to initiate gNodeB UP data transfer, and what role the CP would play in a UP data transfer.  **Disagree**  1. It isn’t clear what is meant by interoperability issues. As stated by QC, one interpretation could refer to compatibility of a model with a device. Please clarify the meaning.  2. It needs to be clarified what control plane messages are needed to facilitate the UP data transfer from the gNodeB.  3. Analysis is expected, but it isn’t a con related to UP data transfer.  4. We do not think that this table is the right place to discuss delta configurations as we haven’t defined what that means.  6. DRBs also transmit data in a robust manner, with automatic adaptation and retransmissions. What is unclear is what is gained by the perceived additional robustness of the CP for what is essentially a file transfer. |
| CMCC | OK for all as starting point | OK for all as starting point |
| Samsung | Ok for all as a starting point for the next round of discussion. | Ok for all as a starting point for the next round of discussion. |
| vivo | Bullet 2 should be for solution 2b/3b | Bullet 5 should be support model switch during mobility is not so straightforward. |
| Lenovo | OK for all | OK for all |
| Spreadtrum | Yes to all for a good start | Ok to all |
| ZTE | Regarding the item 2:  Why Control plane overhead is more serious than the UP plane overhead? To our understanding, the signaling overhead is always there no matter which plain is used. | Item 6:  What is the meaning of the robust? If it is reliability, we do not think it can be a CON for UP transmission, we have redundant PDU session and PDCP duplication for UP data transmission |
| Ericsson | No: - 2. We don’t get it.   OK to the rest as baseline | OK as baseline, but understanding that there is no clarity on how 3b could work. |
| CATT | We do not think the solution 2b and 3b can be discussed in the same table. See comment in Q2.  Yes to bullet 1, 2, 5.  For bullet 3, to use UP solution, we do not think it can be efficient to handle model delivery/transfer during mobility.  For bullet 4, why we need to transfer multiple models simultaneously? | Yes to all bullets. |

## 2.5 Discussion on pros/cons of Solution 4

The pros/cons of Solution 4 are listed in the table below and they are exactly the same as listed in [2]. The numbering is put.

|  |  |
| --- | --- |
| **Solution 4** | Server (e.g. OAM, OTT) can transfer/delivery AI/ML model(s) to UE (e.g. transparent to 3GPP) |
| **Pros** | 1. No 3GPP impacts (can check whether there are 3GPP impacts if OAM based) 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 |
| **Cons** | 1. The latency of model transfer and switching during handover may not be guaranteed 2. There may be inter-operability issues, such as:    1. Different implementations may lead to different model performances and a huge burden of model management (e.g., frequent model activation/deactivation)    2. Massive offline coordination is needed or requires lots of coordinations among vendors, especially for the CSI compression use case 3. DRB transmission is generally less robust than SRB 4. When network cannot control the model transfer/delivery, the transfer of large model may impact important and delay sensitive user data traffic 5. Network can do nothing expect for data collection 6. Not compatible with current mobility procedure   New:  UE needs to store multiple models in advance. |

**Q5: What are the comments on pros/cons of Solution 4?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Comments on Pros** | **Comments on Cons** |
| Qualcomm | Yes –  1 and 2. | No –  1-      Latency can be guaranteed by placing the server at the appropriate place.  2-      We do not think the interoperability issue is associated with whether CP or UP-based method is used. It has more to do with who trains the model. If the model is developed by the UE, the gNB needs to know where the model is applicable irrespective of whether CP/UP is used. Furthermore, we already agreed that the network can use meta info for model control and management. If model ID based LCM is used, the network can be provided with meta info. For functionality based LCM, arguments are not valid.  3-      Can be resolved if the network is made aware of the model delivery.  4-      I believe this is not true because based on the QoS requirement the network can provide 5QI. If the network is not aware, it may provide GBR resources. IF other traffic is delay-sensitive then the network can provide GBR with delay-threshold resources.  5-      Not sure why this is an issue. We are talking about the UE side model. |
| Apple | Yes for all as starting point of next round of discussion  - We don't prefer to repeat the same discussion again in previous post-meeting email discussion.  - Our understanding is that the list is just used for starting point of next round discussion, to facilitate discussion.  - In our understanding, the list is drafted based on majority view, which should be respected. We do not see any necessity to preclude any point at this stage. | Yes for all as starting point of next round of discussion  - We don't prefer to repeat the same discussion again in previous post-meeting email discussion.  - Our understanding is that the list is just used for starting point of next round discussion, to facilitate discussion.  - In our understanding, the list is drafted based on majority view, which should be respected. We do not see any necessity to preclude any point at this stage. |
| Huawei, HiSilicon | Ok for all. | Ok for all. |
| OPPO |  | Comments to bullet 3  We need to understand what is robust |
| Interdigital | If OAM based, then 1 is not correct any more | Similar comment regarding the mobility related issue (#6) as in the previous two questions  Not sure about #5: I mean, we didn’t mention LCM specific aspects in the other solutions, why are we talking about that for this solution only? |
| Intel | Yes to all | Yes to all |
| LGE | Yes to all bullets | Yes to all bullets |
| NTT DOCMO | Yes to all bullets. | Agree with Interdigital about #5  In this discussion, we should focus on model transfer in isolation from LCM.  However, I agree that discussion of the relevance of LCM, such as performance aussurance, is needed in the future. |
| NEC | Yes for 1 and first part of 2. | Yes for 1and 3. No for the remaining. |
| Mediatek | FFS: 1, 2. it may not have impact over air interface, but not sure about other part. E.g., how to make the model visible to gNB if LCM is controlled by gNB. | Yes: 1, 6  FFS: 2, 3, 4, 5 |
| Nokia, Nokia Shanghai Bell | **Agree**  1. Caveat that there may be control signaling associated with making the network aware of the model.  **Disagree**  2. This is a 3GPP impact and it isn’t clear that it is in RAN2 scope how this would be specified. | **Agree -** 4  **Disagree**  1. Discussion about model transfer/delivery and/or model switching during handover has not been defined, so this should not be in the table at this time. Regarding latency, we think that it is out of scope for a non-3GPP model transfer/delivery method.  2. Study is required to determine what, if any, interoperability issues there may be.  3. DRBs also transmit data in a robust manner, with automatic adaptation and retransmissions. What is unclear is what is gained by the perceived additional robustness of the CP for what is essentially a file transfer.  5. Wording is incorrect and disagree. The role of the network is FFS.  6. Disagree. Clarify what is meant by the mobility procedure. |
| CMCC | OK for all as starting point | OK for all as starting point |
| Samsung | Ok for all as a starting point for the next round of discussion. | Ok for all as a starting point for the next round of discussion. |
| vivo | Ok for all as starting point | Corresponding to the cons in Solution 1a, we prefer to add: UE needs to store multiple models in advance. |
| Lenovo | OK for all | OK for all |
| Spreadtrum | Yes to all | This method is complex to model LCM. |
| ZTE | Agree with Interdigital. If OAM is set, it may cause the 3GPP impact. | Don’t agree with robust things.  Don’t agree it will cause the data collision with the delay sensitive data transmission, we have 5QI and QoS control  Don’t agree with gNB can do nothing, if OAM is set by NW,NW can do a lot of things. |
| Ericsson | No: - 1. We aren’t 100% sure. So better to remove. | No: - 1. We haven’t seen any conclusion on this.   * 5. LCM related procedures could still be handled. * 6. Similar to 1. Not sure how we can draw such conclusion. |
| CATT | Yes to all bullets. | Yes to all bullets. |

# 3 Conclusion

Based on companies’ comments, some revisions/highlights are directly made in relevant tables. For the bullets highlighted in yellow, they are FFS due to concerns from some companies, while others can be agreeable.

In summary, the agreeable pros/cons of the listed solutions are listed below (the numbering is the same as used in tables in section 2). Other pros/cons are FFS and can be further discussed.

**Proposal 1: Agree on the following expected pros and cons of the listed solutions as a starting point. Other pros/cons are FFS.**

|  |  |  |
| --- | --- | --- |
|  | **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 |

# 4 Reference

[1] R2\_121 Chair Notes 2023-03-01 1900

[2] R2-2301576 Report of [Post120][053][AIML18] model transfer delivery (Huawei) Huawei