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

# 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 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 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 |

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

## 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 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 |
| **Cons** | 1. Face challenges to convey large 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. 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. |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## 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 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] 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) |

**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. |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## 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, 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 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 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## 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 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 |

**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. |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# 3 Conclusion

[To be added]

# 4 Reference

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

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