**SA WG2 Meeting #153E eMeeting S2-220XXXX**

**Elbonia, Oct. 10 – 14, 2022 (revision of S2-20xxxxx)**

**Source: OPPO, Samsung**

**Title: KI#1: Updated evaluation and conclusion**

**Document for: Approval**

**Agenda Item: 9.21**

**Work Item / Release: FS\_AIMLsys / Rel-18**

*Abstract of the contribution: This contribution proposes an update to KI#1 evaluation and a conclusion.*

# 1. Discussion

This contribution proposes an update to KI#1 evaluation and a conclusion.

# 2. Proposal

\*\*\*\*\* Start of Changes \*\*\*\*\*

## 7.1 Key Issue #1: Monitoring of network resource utilization for support of Application AI/ML operations

Six different solutions have been proposed to address aspects of or the complete description of Key Issue #1 in clause 5.1. These solutions as identified in clause 6.0 are #1, #11, #16, #28, #34, and #37. Solutions #16 and #37 being very similar this evaluation only considers solution #37. While the scope of solution #28 is rather broad and apply to general architecture framework of which the discussions should be deferred to the end of the conclusion phase, solutions #1, #11, #34, and #37 are more focused on specific aspects of KI#1.

A set of architecture principles proposed in the context of the above solutions have been identified for KI#1, and this evaluation clause is structured around those identified principles rather than the individual solutions themselves.

The following list shows positively evaluated architecture principles as potential basis for normative work along the corresponding proposing solutions and their relation to the principle:

- **Maximizing the reuse of existing 5GS mechanisms with extensions, when needed, to assist the AF to monitor the 5G network resource utilization**.

- Solution #1 proposes reutilizing the procedures in clauses 4.15.16.6, 4.16.5 and 4.4.2.2 of TS 23.502 [4] to facilitate monitoring of resources that are relevant in the context of AI/ML-based services and applications.

- Solution #11 and Solution #34 propose the existing NWDAF framework in TS 23.288 [6] to be reused for AF resource monitoring purposes in the context of AI/ML-based services and applications as well as specific NWDAF analytics enhancements for application AI/ML traffic status.

- Solution#37 proposes to reuse existing UPF Exposure Service by introducing a new event type "Bit Rate Reporting per QoS Flow" to the existing consumer "NEF" for this service.

- **Support aggregated bit rate monitoring to assist AF to conduct the Application AI/ML operation according to the SLA**.

- Solution #37 propose monitoring the aggregated bit rate among a set of QoS flows of the group of UEs that are part of the FL operation.

- **Support QoS monitoring for AI/ML-based services triggered by AF to assist AF to conduct the application AI/ML operation according to the SLA**.

- Solution #1 proposes the existing QoS monitoring mechanism in TS 23.502 [4] currently applicable to assist URLLC services to be used, and enhanced, in the context of AI/ML-based services and applications. Based on the outcome of the moderated discussions, more discussions and better understandings are needed to justify the new monitoring parameters.

- **New monitoring parameters available at AF**.

- Solution #1 proposes traffic/data volume and session inactivity time as new monitoring parameters to be available at the AF to assist with the operation of the AI/ML application. The monitoring procedures are based on UPF event exposure.

- **AF monitoring of availability of network resources to assist the application in training of its AI/ML algorithms**.

Furthermore, an additional set of architecture principles have also been identified in the context of KI#1 that do not appear suitable towards normative work. Those architecture principles as well as evaluation views for each principle can be found below:

- **Support for AF resource utilization request from NWDAF, PCF, SMF, where the AF can request the network resource utilization related information from NWDAF, PCF and SMF**. It is not clear what is the purpose of AF being aware of which 5GC NF is producing the information, and the principle lacks a description of what type of network resources it is referring to.

**- Network monitoring of resource usage by the operator to ensure there is a balance between the AI/ML traffic and other service traffic.** While the architecture principle holds true in general for 5GS, it is unclear whether this TR and the follow up normative work should develop solutions in line with this principle or it is out of scope.

- **Control plane resource usage monitoring is also required, in addition, to the user plane, to avoid signalling storms**. It does not seem necessary to develop new congestion control mechanisms beyond what is already existing, and it is currently unclear what other control plane resources are meant by the principle.

The following table is to evaluate the applicable solutions in KI#1 with regards to the proposed criteria. Some extra criteria were added on top of what have been discussed during the moderated evaluation in order to provide a more completed considerations when evaluating this KI#1:

Table 7.1-1: Evaluation of KI#1 solutions

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Evaluation Criteria | Solution#1 | Solution#11 | Solution#34 | Solution#37 |
| **1** | Suggest to leverage and/or extending the existing NWDAF analytics to assist AF to monitor the UE's data transmission performance, i.e. impact to NWDAF. | Proposed to extend existing NWDAF analytic mechanisms to monitor DN performance, UE communication & QoS sustainability  The monitoring is done in order to prepare for the upcoming Application AI/ML transmission | Proposed to leverage the AF Influence for traffic routing feature to support the transmission of the Application AI/ML traffic which also triggers "some" NWDAF analytics to monitoring the UE's performance to assist the traffic routing | Proposed to leverage the NWDAF analytic to monitor the existing Application AI/ML data transfer and to report the ongoing performance to the AF | No |
| **2** | Suggest to leverage other mechanism besides NWDAF to obtain UE's performance, i.e. no NWDAF impact. | Proposed to leverage PC, SMF and UDF etc. to monitor UL/DL & round trip delay, traffic/data volume, session inactivity timer, | No | No | Proposed to extend the existing UPF event exposure with new event to report the UE's on-going bit rate periodically to NEF. NEF will then aggregate the group of UEs bit rate and compare it against the Group-MBS threshold to ensure the group performance does not exceed the aggregated throughput performance. If exceeded, NEF will notify the AF for this new event. |
| **3** | Support aggregated bit rate monitoring to assist AF to conduct the Application AI/ML operation according to the SLA.  NOTE: Corresponding to Principle#3 which receives general support during the Moderated AP Discussions. | No | No | No | Yes |
| **4** | Support aggregated performance monitoring reporting for optimization (e.g. minimizing signalling overhead). | No | No | No | Yes |
| **5** | Support QoS monitoring for AI/ML-based services triggered by AF to assist AF to conduct the Application AI/ML operation according to the SLA.  NOTE: Corresponding to Principle#4 which receives general support during the Moderated AP Discussions. | Yes | Yes (Built-in to the AF Influence to Traffic Routing mechanism) | Yes | Partially (focus on bit rate monitoring) |
| **6** | Both the network and the AF have the need to monitor usage of network resources for AI/ML training traffic.  NOTE: Corresponding to Principle#7 which receives general support during the Moderated AP Discussions. | Yes | Yes | Yes | Yes |
| **7** | Assist AF to perform QoS policy modification on the UE after monitoring the UE's performance.  NOTE: Not specific to this KI. | Trigger AF session request to update UE's QoS configuration | Leverage the imbedded procedure of the AF Influence to modify the QoS policy when needed | Not specified | Not specified |
| **8** | Applicable to FL "aggregated" performance monitoring. | No | No | No | Yes |
| **9** | Proposed mechanism (e.g. gating, QoS policy modification, etc.) to be used by AF or by 5GC to respond to the UE's performance monitoring.  NOTE: Not specific to this KI. | No | No | No | No |
| **10** | AF monitoring of availability of network resources to assist the application in training of its AI/ML algorithms. | May be  NOTE: It has dependency on R18 FS\_UPEAS outcome which is still under study. | No  NOTE: Monitoring is focusing on the traffic route condition | Yes  NOTE: Monitoring is focusing on the transmission status (e.g. location, QoS, load, congestion etc.) | No  NOTE: Monitoring is focusing on whether the pre-defined Group-MBR threshold has been exceeded by the group transmission. |
|  | **Pros:** | Able to leverage the existing NWDAF analytics to adjust the QoS policy and to assist the AF's decision (e.g. UE selection) before starting the application AI/ML operation with the proper UE. | Able to reuse existing AF Influence mechanism to monitoring the on-going Application AI/ML operation and to adjust the QoS policy to adapt to the UE's changing condition. | Able to leverage the existing NWDAF analytics to assist the AF to continue monitoring the UE load and QoS condition, network performance etc. during the application AI/ML operation. | Able to reuse the existing UPF exposure mechanism to monitor the bit rate for the selected group of UEs to support aggregated monitoring request from AF.  The monitoring can be applied before or during the Application AI/ML operation.  The mechanism can support FL operation for aggregated bit rate monitoring with minimum signalling overhead between the 5GC and AF. |
|  | **Cons:** | Not always efficient as the monitoring is mainly done on per UE basis to report to the AF, i.e. signalling overheads could be high. | No assistance to the AF on learning the UE's performance before starting the Application AI/ML operation.  Not always efficient as the monitoring is mainly done on per UE basis to collect User Experience and/or DN performance from the AF, i.e. signalling overheads could be high.  Not all performance/resource monitoring are necessarily produced by the NWDAF.  There is no on-going feedback from the 5GC to the AF to monitor each UE's performance for participating in the Application AI/ML operation. | No assistance to the AF on learning the UE's performance before starting the Application AI/ML operation.  Not always efficient as the monitoring is mainly done on per UE basis to report to the AF, i.e. signalling overheads could be high.  Not all performance/resource monitoring are necessarily done by the NWDAF. | Limited to the aggregated bit rate monitoring only. |

\*\*\*\*\* Next Change \*\*\*\*\*

## 8.1 Key Issue #1: Monitoring of network resource utilization for support of Application AI/ML operations

It is proposed that the architecture principles and specific mechanisms below proposed by KI#1 solutions are considered for the normative work on monitoring of network resource utilization for support of application AI/ML operations:

- It is certain that the DL, UL or round trip packet delay measurement would be an essential tool to assist the AF to make assessment for the proper UE(s) to participate in any of the Application AI/ML operations as defined in TS 22.261 [2]. Today 5GS specifications already well define how the existing URLLC services support the AF monitoring on such packet delay measurements. Therefore, existing services supported for URLLC can be re-used for the purpose of packed delay monitoring of AI/ML traffic and operations as described in Solution #1.

- The proposed extensions from Solution#34 for NWDAF-based procedures and data analytics as described in TS 23.288 [6] to collect broader set of input data from various 5GC NFs to enhance the reporting of the Service Experience, User Data Congestion and NF load information analytics are to be supported. They complement the support for Solution#11 which leverages those NWDAF extensions to assist the AF’s decision to determine the best traffic routing support according to the QoS performance requirements for the Application AI/ML traffic.

- When aggregated bit rate monitoring is requested for a set of QoS flows of a group of UEs against the pre-defined Group MBR threshold, the UPF exposure extension proposed by Solution#37 with the new event type ‘Bit Rate Reporting per QoS Flow’ consumed by NEF as described in Solution #37 is a simple useful tool to be supported to monitor the aggregated throughput performance for the specific Application AI/ML operation.

- AF knowledge on the Traffic/data volume and session inactivity time with new monitoring parameters to recognize the current system performance respective to a given UE as described in Solution#1 is to be supported to assist Application AI/ML application operation. The monitoring procedures are based on UPF event exposure as described in Solution #1.

It is understood that training operation for application AI/ML can be assisted by enabling the monitoring of available resources in the AF as above. Further signalling overhead reduction may be needed across these four solutions and how these solutions are organized as the monitoring tools to support the various Application AI/ML operation will be determined during the normative phase.

\*\*\*\*\* End of Changes \*\*\*\*\*