**3GPP TSG-WG SA2 Meeting #153E e-meeting *S2-2208695***

**Elbonia, October 10 – 17, 2022 (revision of S2-220xxxx)**

**Source: Huawei, HiSilicon**

**Title: KI#6: Update Evaluation and Conclusions**

**Document for: Approval**

**Agenda Item: 9.21**

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

*Abstract: This contribution proposes the update of the evaluation and conclusion of KI#6.*

# 1. Introduction/Discussion

According to the architecture assumption, performance KPIs requirements described in clause 7.10 of TS 22.261 need to be taken into account for AI/ML model transfer in 5GS. How to map these KPI requirements into 5GS QoS parameters is discussed under conclusions for KI#6. All the solutions for KI#6 do not propose how to map performance KPIs as specified in clause 7.10 of TS 22.261 into 5GS QoS parameters, it implies all solutions follows existing mechanism to map performance KPIs for AI/ML to 5GS QoS parameters.

**Observation 1: All solutions follow existing mechanisms to map performance KPIs described in clause 7.10 of TS 22.261 to 5GS QoS parameters.**

Binding the AIML traffic to distinct QoS flow from others applications can facilitate the following aspects related to the KI#6.

- Mapping KPI requirements described in clause 7.10 of TS 22.261 to 5GS QoS parameters.

- Monitoring and reporting on performance KPIs described in clause 7.10 of TS 22.261.

- Dedicated QoS Sustainability analytics to assist the AF to prepare for the initial Application AI/ML operation.

**Observation 2: It is beneficial for the related aspects to KI#6 to bind AIML traffic to distinct QoS flow.**

Standardized 5QI values, as defined in the Table 5.7.4-1: *Standardized 5QI to QoS characteristics mapping* in the TS 23.501, are assumed to be frequently used and thus benefit from optimized signalling by using standardized QoS characteristics. According to the KPI requirement for AI/ML operators, it is seen that 5QI value 7 can be applied to Split AI/ML inference for Enhanced media recognition, and 5QI value 10 can be applied to AI/ML model downloading or Federated Learning, as shown below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 5QIValue | Resource Type | Default Priority Level | Packet Delay Budget(NOTE 3) | Packet ErrorRate  | Default Maximum Data Burst Volume(NOTE 2) | DefaultAveraging Window | Example Services |
| 7 | Non-GBR | 70 | 100 ms(NOTE 10,NOTE 13) | 10-3 | N/A | N/A | Voice,Video (Live Streaming)Interactive Gaming**New: Split AI/ML inference for Enhanced media recognition (see TS 22.261)** |
| 10 |  | 90 | 1100ms(NOTE 13)(NOTE 17) | 10-6 | N/A | N/A | Video (Buffered Streaming)TCP-based (e.g. www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.) and any service that can be used over satellite access type with these characteristics**New: AI/ML model downloading or Federated Learning (see TS 22.261)** |

**Observation 3: 5QI#7 can be applied to satisfy the performance requirements of split AI/ML inference for enhanced media recognition and 5QI #10 can be used for AI/ML model downloading or Federated Learning.**

It is noted that there is no standard 5QI which can meet some of the performance KPIs requirements described in clause 7.10 of TS 22.261. For example, the following part of performance KPI requirements cannot be mapped to existing characteristics of a standard 5QI with proper Packet Delay Budget and Default Maximum Data Burst Volume.

|  |  |  |
| --- | --- | --- |
| Uplink KPI | Downlink KPI | **Remarks** |
| Max allowed UL end-to-end latency | Experienced data rate | Payload size | Communication service availability | Reliability | Max allowed DL end-to-end latency | Experienced data rate | Payload size | Reliability |
| 2 ms | 1.08 Gbit/s | **0.27 MByte** | 99.999 % | 99.9 % |  |  |  | 99.999 % | Split AI/ML image recognition |
|  | 4.7 Mbit/s |  |  |  | 12 ms | 320 Mbit/s | **40 kByte** |  | Split control for robotics |

**Observation 5: New standardized 5QIs and corresponding QoS characteristics mapping should be defined for** **Split AI/ML image recognition and Split control for robotics.**

Based on the observations, it is proposed to update the evaluation and add the corresponding principles in the conclusion of KI#6.

# 2. Text Proposal

It is proposed to capture the following changes vs. TR 23.700-80.

\* \* \* \* First change \* \* \* \*

### 7.6.1 QoS performance measurement assistance to Application AI/ML operation

Table 7.6.1-1 lists solutions that addresses how to map performance KPIs into 5GS QoS parameters and the procedure to perform QoS Monitoring for the UE to AI/ML application traffic to perform both the AI/ML split, AI/ML download and federated learning as defined in KI#6.

Table 7.6.1-1: Mapping performance KPIs into QoS parameters. Procedure for Monitoring QoS parameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Solution** | **Covers KI requirements** | **Impacts on NFs** | **Completeness** | **Open issues/Editor´s Note (NOTE 1)** |
| #1 | Covers Delay and Analytics and procedure | Not described yet. | Not complete yet.How the AI/ML server uses analytics is not described. | Whether and how usage report of inactivity time are exposed.NOTE: Need further update based on this meeting solution update |
| #7 | Covers Delay and Bitrate and procedure  | Impacts Nnef\_AFSessiowithQoS (QoS reference is mandatory) and Npcf\_SMPolicyControlImpacts RAN, SMF, UPF, PCF, NEF and AF | Not complete yet.Unclear how the AF request QoS Monitoring, given that the QoS reference is mandatory parameter in Nnef\_AFSessionwithQoS and the QoS parameters in the PCC rule are mandatory. | How to request QoS monitoring only without requesting QoS is performed in not defined.NOTE: an alternative way for AF to subscribe delay and bitrate monitoring is to reuse 4.16.5.2 in TS 23.502, where AF/NEF provides service information to the PCF without other mandatory parameters except UE address and identification of the application session context |
| #15(NOTE 2) | Covers Delay, Bitrate and Reliability. Covers the procedure. | Impacts UE, UPF, SMF, AF.It extends existing PMF with new functionality to monitor and do analytics.Defines new services and service operations for PMF monitoring. | Not complete yet.New services are not described. | Editor's note: Whether and how to support new UE measurement should coordinate with RAN group.Editor's note: How to support the co-existence of ATSSS and AIMLsys with PMF needs further clarification.Open issue: Lack of details to describe how the new NF obtain the QoS monitoring requirements form AF for the target QoS Flow(s) prior to trigger the PMF for the QoS monitoring.  |
| #40(NOTE 2) | Covers the Procedure for QoS Monitoring | Defines a new NF.Impacts are not described yet. | Not complete yet.New services are not described. | QoS Monitoring is activated using PCC Rules that are provided by the PCF to SMF. Whether a new NF can also activate QoS Monitoring needs to be discussed.NOTE: Need further update based on this meeting solution update |
| #42(NOTE 2) | Covers latency, packet loss rate, bandwidth. | Reuses existing QoS Monitoring procedure as part of AF request to determine if the given UE within the group satisfies the new AIML group performance information | The AIML group performance information is defined, including includes Min/Max latency, Min/Max packet loss rate in UL/DL, Duration for the requested QoS, Minimum number of UEs, Max Requested bandwidth DL/UL included in the AIML group information. | Editor's note: It is FFS whether additional parameters are sent from AF as AIML group performance information.How to monitor other parameter than latency is not described.Open Issues: The proposed solution is claimed to be triggered for each iteration of the FL operation. However, no mention of how to unsubscribe/unprovision the set of QoS flows of a group of UEs from previous UE selection cycle. |
| NOTE 1: The evaluation needs updates when the Editor´s Note or open issues are resolved.NOTE 2: These solutions need to be considered to address KI#7 on performance monitoring for a UE or a group of UEs. |

Based on the evaluations above, it can be stated that the AI/ML application request to monitor the latency provides the Requested 5GS Delay for the AI/ML application traffic in the procedure for Setting up an AF session with required QoS procedure as well as the subscription for QoS measurement. This follows the existing procedure for Setting up (or Update) an AF session with required QoS procedure.Monitoring other QoS parameters such as packet loss rate or bandwidth is not described to a level that can be evaluated yet.

Some solutions reuse AF session with required QoS to request and monitor QoS for AI/ML services, there are two solutions that define a new NF to request QoS monitoring, further discussion is required to determine if new NF is needed.

Based on the above, the proposal is that the monitoring and reporting resource utilization is performed for those performance KPIs described in clause 7.10 of TS 22.261 [2], those are Max. allowed UL/DL end to end latency into 5GS Requested latency then provided in the AF request for QoS procedures.

All the solutions map performance KPIs as specified in clause 7.10 of TS 22.261 [2] into 5GS QoS parameters by existing mechanisms. In addition, binding the AIML traffic to distinct QoS flow (i.e. not sharing QoS flow with other applications) can facilitate the following aspects related to KI#6.

- Mapping KPI requirements described in clause 7.10 of TS 22.261 [2] into 5GS QoS parameters.

- Monitoring and reporting resource utilization performed for those performance KPIs described in clause 7.10 of TS 22.261 [2].

- Dedicated QoS Sustainability analytics to assist the AF to prepare for the initial Application AI/ML operation.

\* \* \* \* Second change \* \* \* \*

## 8.6 Key Issue #6: QoS and Policy enhancements

It is recommended to use the following principles as the basis for the normative work.

* The 5GS should bind the AI/ML traffic to distinct QoS flow (i.e. not sharing QoS flow with other applications).
* No unsolvable issue has been identified that justifies new QoS parameter to be defined
* For the QoS monitoring on per UE basis, existing URLLC mechanism as described in Solution#1 should be reused
* For the aggregated throughput and UE’s bit rate monitoring, the conclusion of KI#1 can be applied here to KI#6.

Support for the QoS time window transmission monitoring and negotiation, it is proposed to merge the Solution#38 with Solution#10 by decoupling the aspects of 5GC assistance for the UE member selection as described in Solution#38 and the dependency on RAN for resource allocation as no unresolvable QoS allocation issue was identified. The list of service parameters proposed by Solution#38 to support timing window negotiation for the time dependent QoS need further evaluation during the merge of the two solutions. - Whether existing 5QI values can be applied for AI/ML traffic or new 5QI values should be defined can be decided during normative phase.

\* \* \* \* End of changes \* \* \* \*