**3GPP TSG-SA WG1 Meeting #99e S1-22xxxx**

**Electronic Meeting, xx – xx August 2022** *(revision of S1-22xxxx)*

**Source: OPPO**

**Title: Concerning Reply LS on 5GC information exposure to UE**

**Document Type: Discussion**

**Agenda item: 3 (LSs and related contributions)**

**Document for: Discussion**

**Contact: Yang Xu <xuyang@oppo.com>**

*Abstract: This document considers the issues and proposed resolutions from incoming LS SA2-2205286 and the draft reply S1-220zzz.*

**1. Introduction**

SA2 kindly asks SA1 to provide guidance and confirm any new requirements for network information exposure to UE.

This discussion paper propose the SA1 answer based on AIML related SA1 requirments and concrete technical scenarios.

**2. Discussion**

In TS 22.261 Clause 6.40, the related requirments are as follow.

- Based on operator policy, the 5G system shall be able to provide an indication about a planned change of bitrate, latency, or reliability for a QoS flow to an authorized 3rd party so that the 3rd party AI/ML application is able to adjust the application layer behaviour if time allows. The indication shall provide the anticipated time and location of the change, as well as the target QoS parameters.

- Based on operator policy, 5G system shall be able to provide means to predict and expose predicted network condition changes (i.e. bitrate, latency, reliability) per UE, to an authorized third party.

- Subject to user consent, operator policy and regulatory constraints, the 5G system shall be able to support a mechanism to expose monitoring and status information of an AI-ML session to a 3rd party AI/ML application.

NOTE 2: Such mechanism is needed for AI/ML application to determine an in-time transfer of AI/ML model.

- 5G system shall be able to provide event alerting to an authorized 3rd party, together with a predicted time of the event (e.g., alerting about traffic congestion or UE moving into/out of a different geographical area).

NOTE 3: A 3rd party AI/ML application may use the prediction information to minimize disturbance in the transfer of learning data and AI/ML model data.

**There are some observations indicating the value of 5GC information to UE.**

**Observation-1: AI inference service (between one UE and Application server) is the main use cases which requires . allowing the UE to have direct access to the network data. The fast notification of some network information to UE will be valuable for UE’s adjustment of AI inference.**

As described in TS 22.261 and TR 22.874, most of the AI inference are real time service with quite low E2E latency. The network information exposure to UE helps to guarantee the UE to get the information in short time. This is especially matters a lot in fast-moving case such as a vehicle with auto-driving.

The change of radio situation may change fast, the UE should be notified as soon as possible to adapt the fast change of QoS. This will help to avoid or minimize interruption of the AI inference service.

We also noticed the DP S2-2203768 was mentioning the FL service, however the FL is not the regular usecase to let UE acquire network information directly since the FL is organized by an FL server.

**Observation-2: In many circumstances, only the UE can adjust the application behavious because only UE itself knows its own status (such as the computation capacity/battery status) and the adjustment is based on both the received network information and the UE’s own status.**

The UE usually runs multiple AI services of multiple applications at the same time, and the computing resources utilized by other applications cannot be told to other application servers. Therefore, only the UE can judge its own computing resource utilization so as to complete the adjustment of the AI ​​operation.

As described in TS 22.261 and TR 22.874, the E2E performance of AI inference is determined by both computation and communication. For example, as described in clause 5.1 and 5.5 in TR 22.874, different splitting point is corresponding to different layers to be calculated in UE (computation capacity) and intermediate data rate (UE’s QoS), thus UE receives the network information can be used for the splitting point adjustment so that to avoid the interruption of the E2E latency when QoS is changed.

Table-1: workload and communication requirement for split points (the table 5.5.2-1 in TR 22.874)

|  |  |  |
| --- | --- | --- |
|  | **Approximate output UL data rate (Mbit/s)** | **Computation load in UE** |
| Candidate split point 1 | 120 | Low |
| Candidate split point 2 | 24 | High |

**Observation-3: Allowing the UE to have direct access to the network data will resolve the problem that Application may not have external UE ID for each of all UEs,**

For many 3rd party applications, they may have not configured external UE ID for specific UEs which makes the exposure to AF impossible. However, if the 3rd party can expose the information to UE directly (i.e. “allowing the UE to have direct access to the network data to support its local operation”) this problem would be resolved)

**Observation-4: Alleviate the implementation work due to that application servers are distributed to specific locations**

Edge servers have different serving area and UE may be served based on the location and mobility, connecting all application servers to 5G core will be a huge implementation work, while the direct way to expose the network information to UE can greatly reduce such work load.

**Observation-5: As the AI inference is happening per UE granularity, per UE level information (e.g. a QoS Sustainability Analytics) should be able to be provided to the application in UE while non-UE level information (e.g. NF load analytics) should be provided to AF only.**

The AI inference in UE mainly depends on the per-UE level network information such as the UE based QoS sustainability analytics. No need to expose network or NF level information such as NF load and Network Performance Analytics.

So far, SA1 does not consider the direct access to network data applies to other kind of AIML operations i.e. Federated learning, model distribution.

**4. Proposal**

Provide the following answer to SA2 (captured in LS-S122xxxx).

SA1 find the network information exposure to UE is valuable to realize a good AIML operation performance.

Specifically, AI inference service is the main use cases which requires allowing the UE to have direct access to the network data. The fast notification of some network information to UE will be valuable for UE’s adjustment of AI inference.

SA1 kindly suggest SA2 to allow UE to have a direct access to obtain UE level network information (e.g. a QoS Sustainability Analytics) while non-UE level information (e.g. NF load analytics) is not required.