**3GPP TSG-SA WG1 Meeting #95e S1-213049r09**

**Electronic Meeting, 23 August – 2 Sept. 2021**

|  |
| --- |
| *CR-Form-v12.1* |
| **CHANGE REQUEST** |
|  |
|  | **22.261** | **CR** | **0551** | **rev** | **-** | **Current version:** | **18.3.0** |  |
|  |
| *For* [***HE******LP***](http://www.3gpp.org/3G_Specs/CRs.htm#_blank)*on using this form: comprehensive instructions can be found at* [*http://www.3gpp.org/Change-Requests*](http://www.3gpp.org/Change-Requests)*.* |
|  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Proposed change affects:*** | UICC apps |  | ME | **X** | Radio Access Network | **X** | Core Network | **X** |

|  |
| --- |
|  |
| ***Title:***  | Adding requirements for AMMT |
|  |  |
| ***Source to WG:*** | OPPO |
| ***Source to TSG:*** | SA1 |
|  |  |
| ***Work item code:*** | AMMT |  | ***Date:*** | 2021-08-23 |
|  |  |  |  |  |
| ***Category:*** | **B** |  | ***Release:*** | Rel-18 |
|  | *Use one of the following categories:****F*** *(correction)****A*** *(mirror corresponding to a change in an earlier release)****B*** *(addition of feature),* ***C*** *(functional modification of feature)****D*** *(editorial modification)*Detailed explanations of the above categories canbe found in 3GPP [TR 21.900](http://www.3gpp.org/ftp/Specs/html-info/21900.htm). | *Use one of the following releases:Rel-8 (Release 8)Rel-9 (Release 9)Rel-10 (Release 10)Rel-11 (Release 11)…Rel-15 (Release 15)Rel-16 (Release 16)Rel-17 (Release 17)Rel-18 (Release 18)* |
|  |  |
| ***Reason for change:*** | Based on the outcome of the SA1 Rel-18 FS-AMMT study, new service requirements for AI/ML model transfer in 5G system need to be specified. |
|  |  |
| ***Summary of change:*** | Add the general description of AI/ML operation and service requirements for AI/ML model transfer in 5G system. |
|  |  |
| ***Consequences if not approved:*** | The performance requirements for AI/ML model transfer in 5G system are not addressed.  |
|  |  |
| ***Clauses affected:*** | 6.38 (new) |
|  |  |
|  | **Y** | **N** |  |  |
| ***Other specs*** |  | **X** |  Other core specifications  | TS/TR ... CR ...  |
| ***affected:*** |  | **X** |  Test specifications | TS/TR ... CR ...  |
| ***(show related CRs)*** |  | **X** |  O&M Specifications | TS/TR ... CR ...  |
|  |  |
| ***Other comments:*** |  |
|  |  |
| ***This CR's revision history:*** |  |

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Start of Change \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

## 6.38 AI/ML model transfer in 5GS

### 6.38.1 Description

Artificial Intelligence (AI)/Machine Learning (ML) is being used in a range of application domains across industry sectors. In mobile communications systems, mobile devices (e.g. smartphones, automotive, robots) are increasingly replacing conventional algorithms (e.g. speech recognition, image recognition, video processing) with AI/ML models to enable applications. The 5G system can at least support three types of AI/ML operations:

- AI/ML operation splitting between AI/ML endpoints

The AI/ML operation/model is split into multiple parts according to the current task and environment. The intention is to offload the computation-intensive, energy-intensive parts to network endpoints, whereas leave the privacy-sensitive and delay-sensitive parts at the end device. The device executes the operation/model up to a specific part/layer and then sends the intermediate data to the network endpoint. The network endpoint executes the remaining parts/layers and feeds the inference results back to the device.

- AI/ML model/data distribution and sharing over 5G system

Multi-functional mobile terminals might need to switch the AI/ML model in response to task and environment variations. The condition of adaptive model selection is that the models to be selected are available for the mobile device. However, given the fact that the AI/ML models are becoming increasingly diverse, and with the limited storage resource in a UE, it can be determined to not pre-load all candidate AI/ML models on-board. Online model distribution (i.e. new model downloading) is needed, in which an AI/ML model can be distributed from a NW endpoint to the devices when they need it to adapt to the changed AI/ML tasks and environments. For this purpose, the model performance at the UE needs to be monitored constantly.

- Distributed/Federated Learning over 5G system

The cloud server trains a global model by aggregating local models partially-trained by each end devices. Within each training iteration, a UE performs the training based on the model downloaded from the AI server using the local training data. Then the UE reports the interim training results to the cloud server via 5G UL channels. The server aggregates the interim training results from the UEs and updates the global model. The updated global model is then distributed back to the UEs and the UEs can perform the training for the next iteration.

### 6.38.2 Requirements

Based on operator policy, the 5G system shall be able to provide means to allow an authorized third-party to monitor the resource utilisation of the network service that is associated with the third-party.

NOTE 1: Resource utilization in the preceding requirement refers to measurements relevant to the UE’s performance such as the data throughput provided to the UE.

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 mechanisms 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 means to supply 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.

The 5G system shall be able to expose aggregated QoS parameter values for a group of UEs to an authorized service provider.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*End of Change\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*