**3GPP TSG-SA WG1 Meeting #95e S1-213050r3**

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

|  |
| --- |
| *CR-Form-v12.1* |
| **CHANGE REQUEST** |
|  |
|  | **22.261** | **CR** | **0552** | **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:***  | CR22.261v18.3.0 Adding performance 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:*** | 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;
* AI/ML model/data distribution and sharing over 5G system;
* Distributed/Federated Learning over 5G system.

Supporting AI/ML model transfer in 5G system requires higher levels of communication performance, e.g. very high data rate in combination with low latency and high communication service availability.Therefore, the performance requirements for AI/ML model transfer in 5G system should be specified. |
|  |  |
| ***Summary of change:*** | Add the performance 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:*** | 7.10 (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 the Change---

## 7.10 KPIs for AI/ML model transfer in 5GS

The 5G system shall support split AI/ML inference between UE and Network Server/Application function with performance requirements as given in Table 7.10-1.

Table 7.10-1 KPI Table of split AI/ML inference between UE and Network Server/Application function

|  |  |  |
| --- | --- | --- |
| **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** |
| 2ms | 1.08Gbit/s | 0.27 MByte | 99.999 % | 99.9% |  |  |  | 99.999% | Split AI/ML image recognition |
| 100ms | 1.5Mbit/s |  |  |  | 100ms | 150 Mbit/s | 1.5 MByte /frame |  | Enhanced media recognition |
|  |  | 4.7Mbit/s |  |  | 12ms | 320Mbit/s | 40kByte |  | Split control for robotics |
| NOTE 1: Communication service availability relates to the service interfaces, and reliability relates to a given system entity. One or more retransmissions of network layer packets may take place in order to satisfy the reliability requirement. |

The 5G system shall support AI/ML model downloading with performance requirements as given in Table 7.10-2.

Table 7.10-2 KPI Table of AI/ML model downloading

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Max allowed DL end-to-end latency** | **Experienced data rate****(DL)** | **Model size** | **Communication service availability** | **Reliability** | **User density** | **# of downloaded AI/ML models** | **Remarks** |
| 1s | 1.1Gbit/s | 138MByte | 99.999 % | 99.9% for data transmission of model weight factors; 99.999% for data transmission of model topology |  |  | AI/ML model distribution for image recognition |
| 1s | 640Mbit/s | 80MByte | 99.999 % |  |  |  | AI/ML model distribution for speech recognition |
| 1s | 512Mbit/s (see note 1) | 64MByte  |  |  |  | Parallel download of up to 50 AI/ML models | Real time media editing with on-board AI inference |
| 1s |  | 536MByte |  |  | up to 5000~ 10000/km2 in an urban area |  | AI model management as a Service |
| 1s |  22Mbps | 2.MByte | 99.999 % |  |  |  | AI/ML based Automotive Networked Systems |
|  1s |  |  500MByte |  |  |  |  | Shared AI/ML model monitoring |
| 3s | 450Mbit/s | 170MByte |  |  |  |  | Media quality enhancement |
| NOTE 1: 512Mbit/s concerns AI/ML models having a payload size below 64 MB. TBD for larger payload sizes.NOTE 2: Communication service availability relates to the service interfaces, and reliability relates to a given system entity. One or more retransmissions of network layer packets may take place in order to satisfy the reliability requirement. |

The 5G system shall support Federated Learning between UE and Network Server/Application function with performance requirements as given in Table 7.10-3.

Table 7.10-3 KPI Table of Federated Learning between UE and Network Server/Application function

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Max allowed DL or UL end-to-end latency** | **DL experienced data rate** | **UL experienced data rate** | **DL packet size** | **UL packet size** | **Communication service availability** | **Remarks** |
| 1s | 1.0Gbit/s | 1.0Gbit/s | 132MByte | 132MByte |  | Uncompressed Federated Learning for image recognition |
| 1s | 80.88Mbit/s | 80.88Mbit/s | 10Mbyte | 10Mbyte |  TBD | Compressed Federated Learning for image/video processing |
| 1s | TBD | TBD | 10MByte | 10MByte |  | Data Transfer Disturbance in Multi-agent multi-device ML Operations |

---End of the Change---