|  |
| --- |
| *CR-Form-v12.2* |
| **PSEUDO CHANGE REQUEST** |
|  |
|  | **26.927** | **CR** |  | **rev** |  | **Current version:** | **0.4.0** |  |
|  |
| *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 |  | Core Network | **X** |

|  |
| --- |
|  |
| ***Title:***  | pCR on missing architecture text |
|  |  |
| ***Source to WG:*** | Interdigital Finland Oy |
| ***Source to TSG:*** | S4 |
|  |  |
| ***Work item code:*** | FS\_AI4Media |  | ***Date:*** | 2023-11-07 |
|  |  |  |  |  |
| ***Category:*** |  |  | ***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-16 (Release 16)Rel-17 (Release 17)Rel-18 (Release 18)Rel-19 (Release 19)* |
|  |  |
| Reason for change: | During SA4 #124, A contribution S4-230912 on architecture for the PD was approved, as well as another contribution S4-230830 to move text on the same clause to the TR. The second did not take into account the text of the first. |
|  |  |
| ***Summary of change:*** | Update the missing text and the figure |
|  |  |
| ***Consequences if not approved:*** | Missing architecture parts |
|  |  |
| ***Clauses affected:*** |  |
|  |  |
|  | **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 changes

### 5.3.3 Architecture for AI data delivery over 5G



Figure 5.3.3-1 AI data delivery general architecture

An architecture for AI data delivery over 5GS is shown in figure 5.3.3-1. Depending on the service scenario and/or use case, certain dedicated AI/ML logical subfunctions may be mapped to, or instantiated by 5GMS functions.

The 5G AI data delivery system shown in figure 5.3.3-1 includes the following main functional blocks:

- **5G AI Client** running on the UE contains two subfunctions:

- **AI data Session Handler**: A function on the UE that communicates with the network side 5G AI Application Function (AF) to establish and control the configuration of an AI data session. The function may include:

- *AI capability manager* subfunctions that monitors, shares and/or reports UE capabilities with/to the AI capability manager function of the 5G AI AF. This may be used for the selection of the model for a UE inference or for the selection of the UE model subset part for a split inference topology between the UE and the network.

- **AI Data Handler**: A function on the UE that communicates with the 5G AI Application Server (AS) and the AI data Handler to establish an AI data delivery session. The function contains:

- An AI inference engine, which has the capability to perform the inferencing of received (split) AI models.

- An AI data access and delivery function, which handles the access and delivery of user plane AI/ML data, as well as conventional media data including

- download the AI model data for inference process. This includes instantiating an AI data access client to access and retrieve AI models or AI model subsets from local files or over the network (e.g., by streaming or downloading the model from a remote server). The inference engine may comprise format decapsulation and model decoding functions as well as a runtime engine that executes the model from the memory.

- Access/deliver intermediate data when a inference is split between the UE and the network.

- Encode data to deliver with serialization and/or compression technique Or conversely decode the received data with deserialization or decompression technique

- **5G AI-Aware Application**: An external function controlled by the external 5G AI application provider implementing the AI/ML application logic, which includes triggering the delivery of an AI model to the inference engine and obtaining inference results from the inference engine.

- **5G AI AS(Application Server)**: An Application Server that hosts 5G AI data functions. It includes

- An *AI data access and delivery function*, which handles the access and delivery of user plane AI/ML data, as well as conventional media data as described above.

- An *AI inference engine*, which has the capability to perform the inferencing of (split) AI models.

- **5G AI AF(Application Function)**: An Application Function that provides various control and configuration functions to the AI Data Session Handler on the UE and/or to the AI Application Provider. It may relay or initiate a request for different Policy or Charging Function (PCF) treatment or interact with other network functions via the NEF (Network Exposure Function). The Application function can include for example:

- AI capability manager subfunctions monitors, shares and/or reports Network capabilities with/to the AI capability manager function of the AI data Session Handler. This may be used for the selection of the model for a UE inference or for the selection of the UE model subset part for a split inference topology between the UE and the network.

End of changes