**3GPP TSG-SA WG4 Meeting #127-bis-eS4-240643**

**E-meeting, 8 April - 12 April 2024**

**Source: Samsung Electronics Co., Ltd.**

**Title: [FS\_AI4Media] pCR on general architecture**

**Spec: 3GPP TR 26.927 v0.6.0**

**Agenda item: 9.6**

**Document for: Agreement**

**1. Introduction**

An update to the general architecture.

**2. Reason for Change**

Update to mapping of functions to general architecture, removing AI specific subfunctions from the M5 reference point. During the last SA4 #127 meeting it was commented that whilst the M5 reference point may be used for the configuration and control plane signaling of data sessions, its functions are predominantly generic in providing media delivery features (such as dynamic policy and network assistance). Such existing features may be applicable to AI delivery session management and configuration without the need to define an AI specific subfunction.

**3. Proposal**

It is proposed to agree the following changes to 3GPP TR 26.927 v0.6.0.

\* \* \* First Change \* \* \* \*

### 5.3.4 Architecture and components for AI data delivery over 5G

#### 5.3.4.1 Introduction



Figure 5.3.4-1: AI data delivery general architecture

A generalized 5G Media Delivery architecture supporting AI media functionality is shown in figure 5.3.4-1. Depending on the service scenario and/or use case, certain dedicated AI/ML logical subfunctions may be mapped to, or instantiated by the generalized media architecture functions.

#### 5.3.4.2 Network functions and UE entities

In addition to the media related definitions described in TS 26.501, additional definitions for AI data related functions include:

- **Media Client** running on the UE contains two subfunctions:

- **Media 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:

- *Features* that monitors, shares and/or reports UE capabilities with/to the Media 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.

- **Media Access Function**: A function on the UE that communicates with the Media AS and the Media Session 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.

- **Media-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.

- **Media 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.

- **Media AF(Application Function)**: An Application Function that provides various control and configuration functions to the Media Session Handler on the UE and/or to the Media 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:

- Supporting features such as monitoring, sharing and/or reporting Network capabilities to the Media 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 via the Media Access Function.

\* \* \* End of Changes \* \* \* \*