**-Source: Samsung Electronics Co., Ltd., Interdigital Finland Ory**

**Title: [FS\_AI4Media] Considering AI/ML from the 5GMS architecture**

**Agenda Item: 9.7**

**Document for: Agreement**

# 1 Introduction

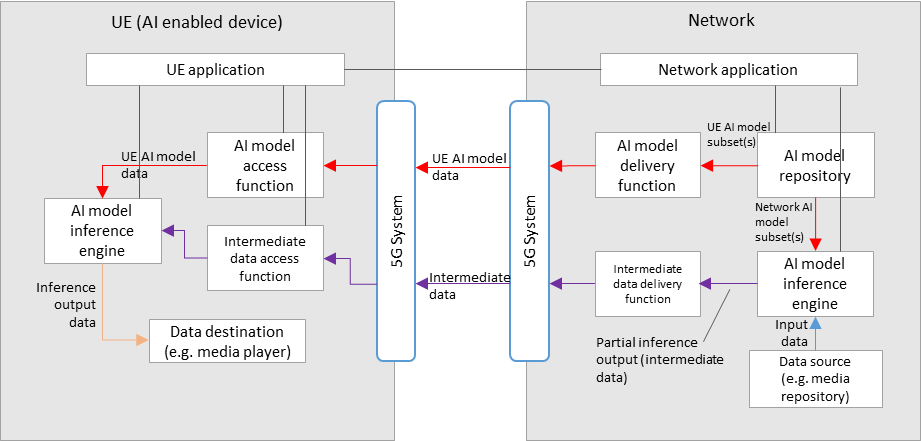
S4-230158r01 is a merge of S4-230158 and S4-230125.

This contribution discusses the possible mapping of AI/ML functions to the 5GMS architecture based on the basic architectures in PD v0.5.

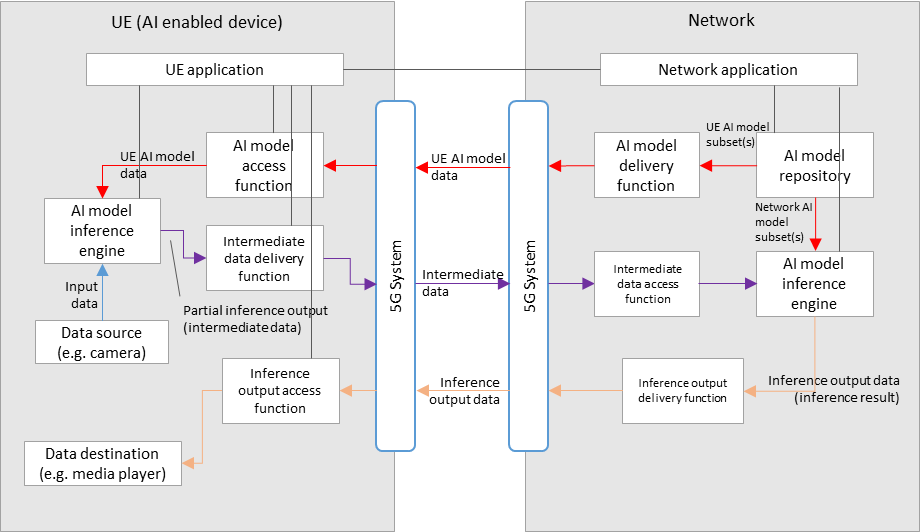
Most AI/ML functions identified in the basic architectures are common to all three scenarios of complete AI/ML model distribution, split AI/ML operation, and distributed/federated learning.

# 2 Architectures

Split AI/ML operation (media source in network)



Split AI/ML operation (media source in UE)



The basic architectures in the PD, as shown in the split AI/ML operation architectures above, show only the user plane data flows and corresponding AI/ML data related logical functions.

AI data components

AI related user plane data include:

* AI model data, including data describing the topology/structure of the AI model, data related to the data nodes of the model, i.e. tensors, and other data which may be dependent on the format used for the AI/Ml model.
* Intermediate data, defined as the output data from the inference process of an AI/Ml model that is not considered the final inference result (depending on the service and output layer of the split AI model, certain intermediate data may have media characteristics, or even be media data). Intermediate data is typically required to be delivered to a second device or entity, as the input to a subsequent second split inference.
* Inference output data, which is the data corresponding to the output result of the final AI inference process for the service. Depending on the nature of the AI data inferencing for the given AI data service, this inference output data may include: labels for identifying recognition like tasks from media, actual media data such as video and/or audio, or perhaps XR related data such as 3D models.

AI4media data logical functions

User plane logical functions supporting the scenarios identified in the PD include:

* AI data delivery function
* AI data access function
* AI model inference engine

For split AI/ML, control plane functions in both the UE and network are needed for configuration, capability exchange and reporting:

* AI capability manager

Mapping functions to 5G architecture for AI data delivery



Figure 1 5G AI data delivery general architecture

A possible architecture for AI data delivery over 5G is shown above. Depending on the service scenario and/or use case, certain dedicated AI/ML logical subfunctions could be mapped to, or instantiated by 5GMS functions.

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

* **5G AI Client** running on the UEcontains 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 the5G 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.
* **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
  + *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.

Example procedure for Split AI/ML operation

The figure below shows an example procedure for split AI/ML operation, including three main parts:

* AI split inference management, and
* AI data delivery session
* Split inference processing



1. Service provisioning and announcement of AI data service on the network side, in particular between the 5GAI AF (application function) and the 5GAI application provider.
2. Service access information acquisition. During this step, the available or required AI model(s) for the service can be made known to the UE, by means of information made available via a URL link pointing to a file or manifest which may list such available AI models. Such additional information may contain AI model specific information, such as the structure, the size, complexity and latency requirements of the AI model.

AI split inference management:

1. Discovering AI data inferencing capabilities and functions in both the UE and network. In this step, the AI capability manger functions in the UE and in the network may use its capabilities to calculate the range of inference latencies for the AI model to be used for the split AI/ML inference service.
2. Requesting AI split inference. Either the UE or the network requests the other side for an AI split inference service. If information describing the AI model was not made known via the service access information in step 2, then such information may also shared during this step.
3. Negotiate splitting the AI inference process. A split point is negotiated between the UE and the network, using information from steps 2, 3 and 4, in order to satisfy the service, capability and AI model inference latency requirements.
4. Acknowledge split and provide the AI data split inferencing access info. In this step, the network (5GAI AF) and UE (AI data session handler) both acknowledge the decided split point, and access information for the AI data is provided to the UE.
5. The split configuration outcome is notified to the 5GAI-aware application.

AI data delivery session

1. Request the start of AI data delivery. On confirmation, the application triggers the 5GAI client to request the start of AI data delivery using the AI data access information provided in step 7.
2. The 5GAI client request the AI data to be deliver from the 5GAI AS.
3. Pipelines for the delivery of AI model data from the 5GAI AS to the 5GAI Client are setup, and suitable delivery sessions are established and initiated. Delivery may be in the manner of streaming delivery, or download delivery (such as that defined in TS 26.501, or any other form of delivery mechanism required by the AI data service.
4. Start inference process in the UE. In this step, the 5GAI client triggers the inference process (the AI inference engine function), namely the UE side of the split inferencing as decided by the result of step 5.
5. Start inference process in the server. In this step, the 5GAI AF triggers the inference process in the 5GAI AS (the AI inference engine function), namely the network side of the split inferencing as decided by the result of step 5.
6. Pipelines for the delivery of intermediate data from the 5GAI AS to the 5GAI Client are setup, and suitable delivery sessions are established and initiated. Delivery may be in the manner of streaming delivery, such as that defined in TS 26.501, or any other form of delivery mechanism required by the AI data service.

Split inference processing

1. The split inference runs between the UE and the network.

# 3 Proposal

We propose to include the figures and descriptions related to the instantiation architecture and procedures into the next version of the permanent document.