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

**Title: [FS\_AI4Media] Updates on procedure for Split AI/ML operation**

**Agenda Item: 9.7**

**Document for: Agreement**

# 1 Introduction

This contribution proposes several updates to the example procedure for Split AI/ML operation in the latest version of the permanent document (v0.6).

# 2 Changes

### 5.3.4 Example procedure for Split AI/ML operation

Figure 5.3.4-1 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. The decision of whether the split point is static or can be updated dynamically during the service may be negotiated. Related metadata may be shared between the network and UE depending on the configuration.
4. Acknowledge the 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. Depending on the specific split inference scenario, the UE and the network may deliver and/or access Intermediate data, Inference output data and/or metadata using the pipelines defined in the AI data delivery session.

Session reporting and update

1. The AI Data Session Handler may collect and send status reports regarding the UE’s AI media service status (for example AI inference status, latency, resource status, capability status, dynamic media properties etc.) to the 5GAI AF.
2. The 5GAI AS may send status reports regarding the network’s AI media service status to the 5GAI AF.
3. The AI Data Session Handler may receive network status, or network AI status reports from the 5GAI AF, as collected in step 16.
4. The AI Data Session Handler may receive media status reports either from the network or internally from the UE.
5. Depending on the configurations negotiated in step 5, as well as related information from the status reports in steps 16, 17 and 18, updates of the AI model selection, split point configuration or the AI data delivery pipelines for the session may take place between the UE and network.

# 3 Proposal

We propose to include the updates in section 2 of this contribution into the next version of the permanent document.