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

**Title: [FS\_AI4Media] High level procedures**

**Agenda Item: 9.7**

**Document for: Agreement**

# 1 Introduction

This contribution is a merge of S4-221378, S4-221418, S4-221420, S4-221467, S4-221468, and S4-221469.

# 2 High level procedures

2.1 Complete/Basic AI/ML model distribution





Steps for the procedures shown:

During the initialization and establishment step, it is assumed that information related to the required features and detailed configurations are exchanged and negotiated between the network and UE. Information may include those related to UE device and network capabilities, AI/ML service information (e.g. service requirements, AI/ML model descriptions), and delivery methods. Such information may be used for the selection of a suitable AI/ML model for the service.

1. The *UE Application* and *Network Application* communicate to trigger AI model delivery, using the information from the initialization and establishment step.
2. An AI model is selected between the *UE Application* and *Network Application*.
3. The *Network Application* identifies the selected AI model in the *AI model Repository/Provider*.
4. The *AI Model Access Function* establishes an AI model delivery session with the *AI Model Delivery Function*.
5. The *AI Model Access Function* receives the AI model.
6. The *AI Model Access Function* passes the AI/ML model to the *AI model Inference Engine* in the UE.
7. Media data from the *UE Data Source* is passed to the *AI Model Inference Engine* in the UE to be used as the input for the AI inferencing.
8. The *AI Model Inference Engine* performs AI inferencing.
9. The *AI Model Inference Engine* passes the inference output result to the *UE Data Destination* for consumption.

2.2 Split AI/ML operation (media source in network)





Steps for the procedures shown:

During the initialization and establishment step, it is assumed that information related to the required features and detailed configurations are exchanged and negotiated between the network and UE. Information may include those related to UE device and network capabilities (including split capabilities), AI/ML service information (e.g. service requirements, split AI/ML model descriptions), and delivery methods. Such information may be used for the selection of a suitable split AI/ML model configuration, and its associated UE and network AI model subsets, for the service.

1. The *UE Application* and *Network Application* communicate to trigger split AI model delivery, using the information from the initialization and establishment step.
2. A split AI model is selected between the *UE Application* and *Network Application*.
3. The *Network Application* identifies the selected UE and network AI model subsets in the *AI model Repository/Provider*.
4. The *AI Model Inference Engine* in the network receives the network AI model subset.
5. The *AI Model Access Function* establishes a UE AI model subset delivery session with the *AI Model Delivery Function*.
6. The *AI Model Access Function* receives the UE AI model subset.
7. In the UE, the *AI Model Access Function* passes the UE AI model subset to the *AI model Inference Engine*.
8. In the network, the *Data Source* passes media data to the *AI model Inference Engine.*
9. The network *AI model Inference Engine* performs network AI inferencing.
10. The *Intermediate Data Access Function* establishes an intermediate data delivery session with the *Intermediate Data Delivery Function*.
11. In the UE, the *Intermediate Data Access Function* receives intermediate data and passes it to the *AI Model Inference Engine*.
12. The *AI Model Inference Engine* in the UEperforms AI inferencing.
13. The *AI Model Inference Engine* passes the inference output result to the *UE Data Destination* for consumption.

2.3 Split AI/ML operation (media source in UE)





Steps for the procedures shown:

During the initialization and establishment step, it is assumed that information related to the required features and detailed configurations are exchanged and negotiated between the network and UE. Information may include those related to UE device and network capabilities (including split capabilities), AI/ML service information (e.g. service requirements, split AI/ML model descriptions), and delivery methods. Such information may be used for the selection of a suitable split AI/ML model configuration, and its associated UE and network AI model subsets, for the service.

1. The *UE Application* and *Network Application* communicate to trigger split AI model delivery, using the information from the initialization and establishment step.
2. A split AI model is selected between the *UE Application* and *Network Application*.
3. The *Network Application* identifies the selected UE and network AI model subsets in the *AI model Repository/Provider*.
4. The *AI Model Inference Engine* in the network receives the network AI model subset.
5. The *AI Model Access Function* establishes a UE AI model subset delivery session with the *AI Model Delivery Function*.
6. The *AI Model Access Function* receives the UE AI model subset.
7. In the UE, the *AI Model Access Function* passes the UE AI model subset to the *AI model Inference Engine*.
8. In the UE, the *Data Source* passes media data to the *AI model Inference Engine.*
9. The UE *AI model Inference Engine* performs UE AI inferencing.
10. The *Intermediate Data Access Function* establishes an intermediate data delivery session with the *Intermediate Data Delivery Function*.
11. In the network, the *Intermediate Data Access Function* receives intermediate data and passes it to the *AI Model Inference Engine*.
12. In the network, the *AI Model Inference Engine* performs network AI inferencing.
13. The UE Data Destination receives the inference output result from the network.

# 3 Proposal

We propose to include the high level procedures related to the corresponding basic architectures into the next version of the permanent document.