**Source: Huawei Technologies Co., Ltd.**

**Title: AI inference negotiation**

**Agenda Item: 9.8**

**Document for: Agreement**

1. Introduction

This contribution provides an updated call flow for split AI/ML inference negotiation.

1. Discussion

The clause 5.2.2 of FS\_AI4Media PD has defined the basic architecture and two work flows for AI/ML split inference media processing between the network and UE.

However, there are some issues in the current workflows for AI/ML split inference:

* The AI/ML split inference negotiation procedure (including the initialisation & establishment step, step 1- Trigger AI model delivery, and step 2 - Select split AI/ML model) is not clear enough to understand how and when the AI/ML model is selected.
* The model delivery steps in the two workflows are duplicated.

Therefore, this contribution proposes to combine the two workflows into one and clarify the split inference negotiation procedure.

1. Proposed changes

--------------------------------------------- Begin changes -----------------------------------------

5.2.2.2 Basic workflows

Figure 5.2.2.2-1 shows a basic workflow for split inference between the network and UE. Steps for the procedures shown are described below.

 

**Figure 5.2.2.2-1: Basic workflow for split inference between the network and UE**

0. The session is established between the UE and the network.

 **AI Split Inference Negotiation** (This step may be performed at the beginning or during the session when the UE or network status has changed):

1. The *UE Application* gets the UE’s capability information, which may include the AI inference processing capabilities, supported AI framework information, connection capabilities, etc.

2a. When the *UE Application* discovers the UE’s local capabilities can’t meet the AI service requirement, it sends an AI split inference request to the *Network Application* with UE’s capability information and the service requirement information.

3a. The *Network Application* gets the network’s capability information, which includes the AI inference processing capabilities, supported AI framework information from the *AI Inference Engine*.

4a. The *Network Application* selects a proper AI model (including the UE AI model subset and the network AI model subset) for split inference from all matched AI models (with different candidate split points) based on the service requirement information, the UE’s capability information and the network’s capability information.

5a. The *Network Application* sends an AI Inference Resource Allocation request to the *AI Model Inference Engine* with the selected network AI model subset information (including the split point and the intermediate data information).

6a. The *AI Model Inference Engine* responds with a successful result to the *Network Application*.

7a. The *Network Application* sends the AI Split Inference Response with the selected UE AI model subset information (including the split point and the intermediate data information) to the *UE Application*.

**Alternative Case#2: UE decides the split inference:**

2b. The *UE Application* sends an AI Model Information Request to the network with the UE’s capability information and the service requirement information*.*

3b. The *Network Application* collects all matched AI models with different candidate split points based on the service requirement information, the UE’s capability information and the network’s capability information.

4b. The *Network Application* sends the AI Model Information Response with all matched UE AI model subset(s) information (including the split point and the intermediate data information) to the *UE Application*.

5b. The *UE Application* selects a proper AI model based on the UE’s capability information and the received information in the AI Model Information Response.

6b. The *UE Application* sends an AI Split Inference Request to the *Network Application* with the selected AI model information.

7b. The *Network Application* sends an AI Inference Resource Allocation request to the *AI Model Inference Engine* with the network model subset information corresponding to the AI model selected by the *UE Application*.

8b. The *AI Model Inference Engine* responds with a successful result to the *Network Application*.

9b. The *Network Application* sends the AI Split Inference Response to the *UE Application*.

**AI Model Subset Delivery:**

10. The *Network Application* identifies the selected UE and network AI model subsets in the *AI model Repository*.

11. The *AI Model Inference Engine* in the network receives the network AI model subset.

12. The *AI Model Access Function* establishes a UE AI model subset delivery session with the *AI Model Delivery Function*.

13. The *AI Model Access Function* receives the UE AI model subset.

14. In the UE, the *AI Model Access Function* passes the UE AI model subset to the *AI model Inference Engine*.

**AI split inference:**

 **Alternative case#1: data source in the network**

15a. The network *AI model Inference Engine* receives media data from the network *Data Source* or a peer user.

16a. The network *AI model Inference Engine* performs network AI inferencing.

17a. The *Intermediate Data Access Function* establishes an intermediate data delivery session with the *Intermediate Data Delivery Function*.

18a. In the UE, the *Intermediate Data Access Function* receives intermediate data and passes it to the *AI Model Inference Engine*.

19a. The *AI Model Inference Engine* in the UE performs AI inferencing.

20a. The *AI Model Inference Engine* passes the inference output result to the *UE Data Destination* for consumption.

**Alternative case#2: data source in the UE**

15b. In the UE, the *Data Source* passes media data to the *AI model Inference Engine*.

16b. The UE *AI model Inference Engine* performs UE AI inferencing.

17b. The *Intermediate Data Access Function* establishes an intermediate data delivery session with the *Intermediate Data Delivery Function*.

18b. In the network, the *Intermediate Data Access Function* receives intermediate data and passes it to the *AI Model Inference Engine*.

19b. In the network, the *AI Model Inference Engine* performs network AI inferencing.

20b. The network *AI Model Inference Engine* sends the inference output result to the *UE Data Destination* or a peer user.

--------------------------------------------- End changes -----------------------------------------

1. Proposal

We propose to update the clause §5.2.2.2 of the permanent document with the proposed changes.