**3GPP TSG-SA WG4 Meeting #128S4-241029\_r01**

**Jeju, South Korea, 20 May - 24 April 2024**

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

**Title: [FS\_AI4Media] pCR on fix of procedure for split inferencing**

**Agenda item: 9.6**

**Document for: Agreement**

**1. Introduction**

The contribution provides a revised procedure figure for split inferencing.

**2. Discussion**

The current figure 5.3.5-1 in TR 26.9.27 v0.7.0 incorrectly specifies the provision step in step 1 as between the Media AS and the Media Application Provider. This step is corrected to that of between the Media AF and the Media Application Provider

**3. Proposal**

It is proposed to agree the following changes to 3GPP TR 26.927 v0.7.0.

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

### 5.3.5 Procedure for Split AI/ML operation

Figure 5.3.5-1 shows a procedure for split AI/ML operation, including three main parts:

- AI split inference management, and

- AI data delivery session

- Split inference processing



Figure 5.3.5-1: Procedures for split AI/ML operation

1. Service provisioning and announcement of AI data service on the network side, in particular between the Media AF (application function) and the Media 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 a list of features available from each AI model, including its variants, including specific information such as the inferencing accuracy, the size, the amount of nodes, structure, complexity and latency requirements of the AI model.

AI split inference management:

3. 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

4. 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 be shared during this step.

5. 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 whether it can be updated dynamically during the service may be negotiated. Related metadata may be shared between the network and UE depending on the configuration and a set of split points can be negotiated.

6. Acknowledge the split and provide the AI data split inferencing access info. In this step, the network (Media 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.

7. The split configuration outcome is notified to the Media-aware application.

Split AI data session

8. Request the start of intermediate data delivery. On confirmation, the application triggers the Media Client to request the start of AI data delivery using the AI intermediate data access information provided in step 7.

9. The Media client request the intermediate data to be delivered from the Media AS.

10. Pipelines for the delivery of AI model data from the Media AS to the Media 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.

11. Start inference process in the UE. In this step, the Media 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.

12. Start inference process in the server. In this step, the Media 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.

13. Pipelines for the delivery of intermediate data from the Media AS to the Media 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

14. 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.

Split point update and inference processing

15. A split point update is triggered, for example from the media aware application to adapt to the new conditions (e.g. UE capabilities or network capacity has changed). The new split point metadata information is either negotiated between the UE and the network or pass alongside the delivery pipeline from the UE to the network side.

Session reporting and update

16. The Media 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.

17. The Media AS may send status reports regarding the network’s AI media service status to the Media AF.

18. The Media Session Handler may receive network status, or network AI status reports from the Media AF, as collected in step 16.

19. The Media Session Handler may receive media status reports either from the network or internally from the UE.

20. 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.

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