**Source: Interdigital Finland Oy**

**Title: [FS\_AI4Media] Conclusions for split operations**

**Agenda item: 9.6**

**Document for: Agreement**

1. **Introduction**

TR 26.927 started more than 2 years ago as part of the feasibility study on artificial intelligence and machine learning for media (FS\_AI4Media) with the aim to identify and study the following key operations applied to 5G media services

1. AI/ML operation splitting between AI/ML endpoints
2. AI/ML model/data distribution and sharing over 5G system
3. Distributed/Federated Learning over 5G system

The output of an offline meeting held on June 24th invited proponents to start writing the conclusion for preparing a Stage 2 procedures for key operations on AI media data.

This contribution proposes a conclusion clause for the splitting operations

1. **Reason for Change**

The technical report would not be complete, and it will not be possible to adopt agreements of this study into any stage-2 or stage-3 specifications.

1. **Proposal**

It is proposed to agree the following changes to the functional Permanent Document v1.3.0 of the TR 26.927. The proposal is an initial version of the conclusion for discussion.

\* \* \* Begin Changes \* \* \* \*

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

# 10 Conclusion

The objectives of this study were primarily to identify relevant interoperability requirements and implementation constraints of AI/ML in 5G media services applied to model delivery, split operations, and federated learning based on the initial considerations in TS 22.261.

## 10.1 Split operations between AI/ML endpoints

In this study, an architecture and procedures for enabling split operations between a first endpoint and a second endpoint were defined, where an endpoint is either in the UE or in a network server. The study includes:

* High level architecture and procedures developed before being mapped to 3GPP services (5GMS, RTC [and IMS])
* Procedures for the selection of a model and a split point configuration in both decision modes, UE or network centric.
* Mapping AI functions to the generalized 5G media delivery architecture addressing underlying 5GMS, RTC [and IMS] architecture.
* Procedures and call flows for a complete split operation including split inference management, data delivery session for model data and intermediate data, and split inference processing and session reporting.
* Description of metadata negotiated and exchanged between endpoints for identified 3GPP services. These metadata can be exchanged via the control plane or the user-plane. They include:
	+ Model information common to the three key operations
	+ Service requirements common to the three key operations
	+ Endpoint capabilities common to the three key operations
	+ Intermediate data information for split operations
	+ Compression settings for intermediate data

In addition, evaluations have been performed to assess the feasibility of model splitting, split model inferencing and intermediate data compression:

* Model splitting applied to a large set of pre-trained ONNX models, with splitting at any node in the single or multi-branch model, multi-branch beyond 100 branches.
* Split model Inferencing for some of these different split point configurations.
* Different techniques for intermediate data compression offer different capabilities: from model agnostic to model specific, with or without retraining, support for single or multiple split points. Results for these different techniques provides evidence of near lossless compression with a limited impact on the accuracy. Compression ratios of these techniques vary widely depending on the chosen capabilities/requirements.

[Based on the details in this report, the next steps proposed are :

* Document the stage 2 procedures for split operations based on the generalized 5G architecture in TS 26.501 for 5GMS and TS 26.506 for RTC.
* Study options for supporting intermediate data compression.
* Document the relevant split operations stage 3 procedures.

Note: identification of the release for the normative work is TBD]

## 10.2 Model/data distribution

## 10.3 Distributed/Federated Learning

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