**Source: Interdigital Finland Oy**

**Title: [FS\_AI4Media] pCR on conclusions for split operations**

**Spec: 3GPP TR 26.927 v0.8.0**

**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 conclusions 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 3GPP TR 26.927 v.0.8.0

\* \* \* first Change \* \* \* \*

# 2 References

\* \* \* End of first change \* \* \* \*

\* \* \* second Change \* \* \* \*

# 10 Conclusions

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)
* 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 or RTC 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 which can be carried out in the control plane or the user-plane and includes:
  + 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 to this report, an evaluation has been launched to assess the feasibility of the split operations listed above. Evaluations have been performed for 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 different split point configurations above.

Intermediate data compression with existing tools applied to different split point configurations showed almost 1:5 to 1:10 of data size reduction with a limited impact on the accuracy.

[ Editor’s note: The current draft conclusion is based on the functional TR 26.927 v0.8.0 and the Permanent document draft v0.6.0 of the evaluation TR 26.847]

Besides, the MPEG FCM (Feature Compression for Machine) dedicated group for feature/intermediate data compression provides early near lossless results with a reduction of intermediate data size ranging from 6000:1 to 40000:1 and 75% overall bitrate reduction over the same range of task performance as the remote inferencing anchor for video tasks.

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

In the short-term:

- Document the relevant split operations stage 2 procedures based on the generalized 5G architecture in TS 26.506 as addressed in TR 26.927 clause §5.3.

In the mid-term:

* Study options for supporting intermediate data compression such as FCM

- Document the relevant split operations stage 3 procedures based on the generalized 5G architecture in TS 26.506 as addressed in TR 26.927 clause §5.3.

## 10.2 Model/data distribution

## 10.3 Distributed/Federated Learning

\* \* \* End of second changes \* \* \* \*