**3GPP TSG-SA WG4 Meeting #127-bis-eS4-240641\_r01**

**E-meeting, 8 April - 12 April 2024**

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

**Title: [FS\_AI4Media] On operator identified challenges and proposals**

**Agenda item: 9.6**

**Document for: Discussion and agreement**

**1. Introduction**

During the last SA4 #127 meeting in Sophia-Antipolis, the group received input on operator identified challenges and proposals for AI/ML. The text related to these challenges and proposals were included into the functional PD v1.1.0. This contribution aims to trigger further discussions on clarifications and possible requirements which may be extracted from the text.

**2. Discussion**

Comments for discussion on each of the identified challenges/proposals are given below:

* given the recent technical advancements in client devices and all scenarios included in the study, to re-evaluate the trade-off between having powerful client devices capable of executing complex AI workloads and offloading heavy computational task from clients to the network (edge)

EY: our study already considers the requirement of split inferencing and the impacts of different split points on processing complexity for both the UE and network, but complete offloading as a service (not just a configuration) has not been considered. If interest is identified by the group remote AI offloading should be added into our study.

* to test more challenging network scenarios including eNB handovers, client device using different mobile operators, international roaming, switching between different operational modes: standalone operation of the client devices versus co-operative operation with the application part on the MEC/edge node when available.

EY: seems to be out of SA4 scope, but standalone vs co-operative with MEC/edge is related to split inferencing - such configuration is something that should be in scope of SA4 (considering M4)

* to evaluate the performance (in terms of bandwidth and latency) of 5G and “beyond 5G” networks

EY: out of SA4 scope.

* to consider the scalability of the architecture with increasing number of client devices simultaneously making use of the same application against the same MEC node. Main concerns, especially with scenarios in Section 5.1.1.1 in TR 26.927, are limited uplink bandwidth and resource sharing at the edge.

EY: uplink requirements are related to the extraction of uplink data from our use cases, such as intermediate data and possibly inference results and media data (in the case of complete offloading). Without further details it is difficult to extract what SA4 can consider regarding resource sharing at the edge.

* to assess the complexity of modern popular AI/ML models (like EfficientNet, EfficientDet, and YOLO) that may not be suitable for a clear split/partition like the old model used in the Vodafone research study.

EY: the evaluation work started has touched on this issue. What would be a clear requirement from this assessment? It would depend on the service provider on whether to only support clear split models, or whether to also support the split of more complex models. This may also be related to interoperability considerations on formats of AI models for our study.

* to consider alternative approaches like “ensemble models”: complementing a simple, fast in-vehicle AI/ML model with more sophisticated additional AI/ML models executed on the MEC node to increase confidence scores for detection results.

EY: highly relevant to state of the art LLM and LAM use cases in the industry. Whether ensemble model approaches can also be considered a type of split inferencing needs to be studied. Some ensemble model use cases chain independent models for inference. This may also be already related to some of our use cases in the study. The group should discuss on whether to consider this as an extra (split) scenario).

**2. Proposal**

Through the discussions during the meeting it was suggested to include a list of operator topics of interest which may be used as a comparison with the work and objectives being addressed in the study item. If necessary, the update or addition of objectives to the study item may also be considered.

It is proposed to include the following list into the functional PD as an interpretation of the bullet points as operator interest topics (related to SA4):

* Split inferencing between clients and the network, in particular trade-offs on different split configurations and the impacts on processing capability.
* Impacts of AI/ML media services on uplink requirements, in particular limited uplink bandwidth and resource sharing at the edge, where an increasing number of client devices may simultaneously make use fo the same application against the same MEC node.
* Exploring the complexity, and in particular the splitability of the modern state of the art popular AI/ML models for split inferencing.
* Exploring alternative approaches to split inferencing, such as “ensemble models”, where independent AI models with different complexity requirements may be inferenced in the UE and the network separately, but used together for the same AI media service.