**Source:** Nokia Corporation1

**Title:** [FS\_AI4Media] Progressive model delivery call flows

**Document For:** Agreement

**Agenda item:** 9.6

# Introduction

In SA4 #124, a new scenario description on bit-incremental transmission and deployment of AI/ML models was added to the PD. The use case describes a scenario where a model is first downloaded at a low-precision to enable fast startup of inference at the UE. Later the model is updated to a higher/full-precision version with a model update to accrue the benefits of higher/full-precision model inference seamlessly.

Machine learning models are growing in complexity and size. Even with model compression, the transmission size and hence time may be high. In a typical model delivery scenario, the UE may start inference operations only once the model is fully downloaded, uncompressed and deployed. With bit incremental model delivery, a lower-precision but smaller transmission size model is delivered to the UE first. This enables the UE to start inference quickly as compared to the case where it has to wait for a full precision model to be downloaded.

Evaluation of the scenario is ongoing; as a first step a scenario evaluation description ([S4aV230072](https://www.3gpp.org/ftp/TSG_SA/WG4_CODEC/3GPP_SA4_AHOC_MTGs/SA4_VIDEO/Docs/S4aV230072.zip)) was presented and agreed in SA4-e (AH) Video SWG post 125e. Bit Incremental model delivery is a subset of a potentially broader set of progressive model delivery techniques, wherein a low precision model is delivered to a UE first to expedite inference at the UE and to improve QoE. Subsequent model updates are delivered to the UE and the model at the UE is updated to higher precision.

In this contribution, we propose a basic call flow for progressive model deployment based on the model deployment call flow described in clause 5.2.1.2.1 in the PD.

# Progressive Model Delivery



Progressive model delivery refers to a model delivery paradigm wherein a low precision model is delivered to a UE first to expedite inference at the UE and to improve QoE. Subsequent model updates are delivered to the UE and the model at the UE is updated to higher precision. The update may be applied to the model in different ways, depending on how the low precision model is obtained. For example, in case of bit-incremental model delivery the model update may be applied in an additive manner.

* 1. The UE Application and Network Application communicate to establish a progressive model delivery session. The UE Application may receive Service Access information to learn about available services and configurations, including available models, precisions and possible updates. This information may be in a 3GPP URI of/or model manifest file(s). The model manifest file contains size, complexity information etc. of the different versions.
  2. An AI model is selected by the UE Application, based on, e.g. model size and currently available network capacity.
  3. The UE application requests the selected model from the Network Application
  4. The Network Application identifies the selected AI model in the AI model Repository/Provider.
  5. The AI Model Access Function establishes an AI model delivery session with the AI Model Delivery Function.
  6. The AI Model Access Function receives the AI model of the precision requested by the UE.
  7. The AI Model Access Function passes the AI/ML model to the AI model Inference Engine in the UE.
  8. Inference loop: The Data Source passes data to the AI model Inference Engine, AI Model Inference Engine performs AI inferencing, and AI Model Inference Engine passes the inference output result to the UE Data Destination for consumption.
  9. The UE application triggers a model precision update (parallel to the inferencing loop of step 8). The update is a precision update of the model currently at the UE (steps 6-7) rather than a new model.
  10. The model update is delivered to the AI model access function
  11. The model in the inference engine is updated to a higher precision using the model update from 10.
  12. Steps 9-11 may be repeated as 12-13 depending upon number of precision levels and corresponding model updates

# Proposal

We propose the call flow in section 2 be added to the PD.