**3GPP TSG-SA5 Meeting #158 *S5-247092***

Orlando, USA, 18 - 22 November 2024

**Source: Nokia, ZTE, Ericsson?, NTT DOCOMO**

**Title: Rel-19 pCR TR 28.858 Add evaluation for sustainable AIML**

**Document for: Approval**

**Agenda Item: 6.19.1**

# 1 Decision/action requested

***The group is asked to discuss and agree on the proposal.***

# 2 References

[1] 3GPP TR 28.858 “Study on Artificial Intelligence / Machine Learning (AI/ML) management Phase 2”.

# 3 Rationale

This pCR is to add an evaluation for the already agreed use cases and solutions on sustainable AI/ML during previous meetings.

# 4 Detailed proposal

5.1.8 Sustainable AI/ML

5.1.8.1 Description

Sustainable AI/ML aims to reduce energy usage in the ML model training step. It is important to investigate effective methods for evaluating AI/ML energy consumption and efficiency for ML model training which will aid in decreasing energy consumption, enhancing energy efficiency, and promoting sustainable development of AI/ML in 5G systems.

5.1.8.2 Use cases

5.1.8.2.1 AI/ML energy consumption evaluation and reporting for ML model training

AI/ML energy consumption is a significant concern, especially as ML models become more complex and data-intensive. The ML model training is typically the most energy-intensive part since it involves processing large datasets and performing numerous calculations to adjust the ML model parameters.

The MnS consumer may have concern on the energy consumption for ML model training and the MnS Producer should report the related information. MnS Consumer may ask the MnS Producer to report the energy consumption of training each ML model. To satisfy such query request from the MnS consumer, methods should be investigated to measure AIML energy consumption. Then, MnS Consumer can use the energy consumption information and compare with the network performance gain due to ML model to evaluate the total benefit of using the Ml model. Alternatively, the MnS consumer may ask the MnS Producer to report the number of floating-point operations for training each ML model.

Regarding the various AI/ML energy consumption aspects outlined in clause 4.X, ML model training may be characterized by several factors including:

- Number of cycles, a low number of cycles generally contributes to reduced energy consumption,

- Dataset size, large datasets generally contribute to increased energy consumption due to intensive data processing,

- Number of floating-point operations, complex ML model contributes to increased energy consumption,

- Deployment on high-end hardware platform, which can either increase energy consumption due to

their substantial power requirements, or reduce it by offering higher energy efficiency through

optimized performance.

5.1.8.3 Potential Requirements

REQ-AIML\_TRAIN\_ENERGY\_MGT-01: The MLT MnS producer should have a capability to allow an authorized MnS consumer to query the number of floating-point operations pertaining to training each ML model.

REQ-AIML\_TRAIN\_ENERGY\_MGT-02: The MLT MnS producer should have a capability to report the number of floating-point operations pertaining to training each ML model to an authorized consumer.

5.1.8.4 Possible solutions

Solution 1: The IoC MLTrainingReport needs to be enhanced with a new attribute related to energy consumption for training each ML model, e.g., modelEnergyConsumptionTraining. This attribute is of type real and indicates the energy consumption value pertaining to the ML model training, e.g., “KWh”.

Editor’s Note: The method to determine energy consumption for ML model training is FFS. If energy consumption for ML model training cannot be determined, the proposed solution 1 in 5.1.8.4 becomes redundant.

Solution 2: The IoC MLTrainingReport needs to be enhanced with a new attribute, e.g., flopsTraining, to indicate the number of floating-point operations needed to train an ML model. This attribute is of type integer.

5.1.8.5 Evaluation

Solution 1 proposes adding an attribute to the IOC MLTrainingReport that indicates the energy consumption for training an ML model. There is also a note that the method to determine energy consumption is FFS and if this cannot be determined the solution becomes redundant. The method for determining energy consumption has not been determined as part of this TR, therefore no recommendations are made for this solution to be directly used in normative phase.

Solution 2 described in clause 5.1.8.4 proposes the addition of a new attribute to the MLTrainingReport IoC to enable the MnS producer to report the number of floating-point operations needed to train an ML model which is a feasible solution to be developed further in the normative specifications. While the number of FLOPs represent one aspect of energy consumption during ML model training, other aspects may contribute to the overall energy consumption that needs to be considered in the future. Furthermore, it's important to note that this solution may not be directly applicable to pre-trained ML models.

***Start of next change.***

### 5.5.2 Sustainable AI/ML

#### 5.5.2.1 Description

AI/ML techniques have been integrated into mobile networks enabling the learning of network states and provide informed recommendations on the most effective actions based on these learnings. This approach significantly enhances the efficiency in managing a wide array of network use cases. These include, but are not limited to, mobility management, load balancing, and resource optimization. The expectation is that the application of AI/ML techniques will lead to an improvement in overall network performance due to the ability of these techniques to provide more accurate predictions and more effective decision-making processes. However, it is important to note that while the use of AI/ML techniques can lead to significant improvements in network performance, these benefits may come at the cost of increased energy consumption. AI/ML-based solutions, due to their complex computational requirements, may consume additional energy compared to traditional, non-AI/ML solutions.

Sustainable AI/ML aims to reduce energy usage in the AI/ML inference step. It is important to investigate effective methods for evaluating AI/ML energy consumption and efficiency for AI/ML inference which will aid in decreasing energy consumption, enhancing energy efficiency, and promoting sustainable development of AI/ML in 5G systems.

#### 5.5.2.2 Use cases

##### 5.5.2.2.1 AI/ML energy consumption evaluation and reporting for AI/ML inference

AI/ML energy consumption is a significant concern, especially as ML models become more complex and data-intensive. The ML model once trained is deployed for inference which consumes a considerable amount of energy. Each inference made by the ML model requires computational resources, and these resources in turn require energy to operate. With potentially thousands of inferences being made, the total energy consumption during inference can be substantial. Alternatively, the MnS consumer may ask the MnS Producer to report the number of floating-point operations for generating AI/ML inferences.

The MnS consumer may have concern on the energy consumption for AI/ML inference and the MnS Producer should report the related information. MnS Consumer may ask the MnS Producer to report the energy consumption for inference. To satisfy such query request from the MnS consumer, methods should be investigated to measure AIML energy consumption during the inference. Then, MnS Consumer can use the energy consumption information and compare with the network performance gain due to ML model to evaluate the total benefit of using the ML model.

#### 5.5.2.3 Potential requirements

**REQ-AIML\_INF\_ENERGY\_MGT-01**: The AI/ML inference MnS producer should have a capability to allow an authorized MnS consumer to query floating-point operations pertaining to generating inferences.

**REQ-AIML\_INF\_ENERGY\_MGT-02**: The AI/ML inference MnS producer should have a capability to report floating-point operations pertaining to generating inferences to an authorized consumer.

#### 5.5.2.4 Possible solutions

Solution 1: The inferenceOutputs attribute (of type inferenceOutput) defined in the AIMLInferenceReport IoC needs to be enhanced with a new attribute related to energy consumption for AI/ML inference, e.g., inferenceEnergyConsumption. This attribute is of type real and indicates the energy consumption value pertaining to generating AI/ML inference, e.g., “KWh”.

Editor’s Note: The method to determine energy consumption for AI/ML inference is FFS. If energy consumption for AI/ML inference cannot be determined, the proposed solution 1 in 5.5.2.4 becomes redundant.

Solution 2: The IoC MLTrainingReport needs to be enhanced with a new attribute, e.g., flopsInference, to indicate the number of floating-point operations needed to generate AI/ML inferences. This attribute is of type integer.

5.5.2.5 Evaluation

Solution 1 proposes adding an attribute to the IOC AIMLInferenceReport that indicates the energy consumption for generating an AI/ML inference. There is also a note that the method to determine energy consumption is FFS and if this cannot be determined the solution becomes redundant. The method for determining energy consumption has not been determined as part of this TR, therefore no recommendations are made for this solution to be directly used in normative phase.

Solution 2 described in clause 5.5.2.4, proposes the addition of a new attribute to the AIMLInferenceReport IoC to enable the MnS producer to report the number of floating-point operations needed to generate AI/ML inference which is a feasible solution to be developed further in the normative specifications. While the number of FLOPs represent one aspect of energy consumption during AI/ML inference, other aspects may contribute to the overall energy consumption that needs to be considered in the future. Furthermore, it's important to note that this solution may not be directly applicable to pre-trained ML models.

***End of changes.***