**3GPP TSG-SA5 Meeting #154 *S5-242141***

Changsha, China, 15 April - 19 April 2024

**Source: ZTE Corporation**

**Title: Rel-19 pCR TR 28.908 add use case, requirements and solution for energy consumption control**

**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 TS 28.105: "Management and orchestration; Artificial Intelligence/Machine Learning (AI/ML) management".

[2] SP-231722, “Study on AIML management - phase 2”

# 3 Rationale

The integration of sustainable AI/ML operations within 5G networks represents a forward-looking approach to addressing the environmental challenges associated with the rapidly expanding use of AI technologies in telecommunications. We should find a way to address energy consumption issue associated with AI/ML supported features.

This contribution is related to WT-5 of the FS\_AIML\_MGT\_Ph2 SID [2], which is to investigate the sustainability aspect of AI/ML.

# 4 Detailed proposal

***Start of First change***

# 5 Use cases

## 5.3 Common management capabilities for ML training and AI/ML inference phase

## 5.3.x Sustainable AI/ML Operation

5.3.x.1 Description

Sustainable AI/ML operation is an emerging field that balances the rapid growth of AI technologies with the urgent need for environmental conservation. Sustainable AI/ML operation refers to minimizing the energy consumption in the AIML operation workflow including training, emulation, deployment, and inference phases. The energy consumption may be very different in different phase depending on the complexity of the task and selected MLEntity.

Therefore, it is urgent to study the effective evaluation and control of energy consumption and energy efficiency of AI/ML technology, so as to help reduce energy consumption costs, improve energy efficiency, and promote the sustainable development of AI/ML technology.

## 5.3.x.2 Use Cases

5.3.x.2.1 Energy consumption/efficiency control

Energy consumption in MLEntity training and inference is a significant concern, especially as models become more complex and data-intensive. The training phase is typically the most energy-intensive part in the AI/ML operation workflow. It involves processing large datasets and performing numerous calculations to adjust the model's parameters. For inference phase, the energy consumption/efficiency depends on the selected MLEntity. Although the inference phase consumes less energy per task compared with training phase, inference task is performed more frequently and at a larger scale, which may lead to a significant overall energy consumption/efficiency.

In R18, MLTrainingRequest IOC and MLTrainingReport IOC are defined, but the impact of energy consumption/efficiency ise not considered. MnS Consumer may have requirements/policy on the energy consumption/efficiency of the ML training/inference task, the MnS Producer should orchestrate a training/inference solution that meets the energy consumption/efficiency requirements. Furthermore, the energy consumption/efficiency information of each MLEntity should be provided to producer as the reference of MLEntity(s) selection and solution orchestration.

5.3.x.2.2 Energy consumption/efficiency reporting

Performance of AIML needs to be measured for all the aspects of ML entities in different model lifecycle phases including model training, testing, deployment or update and inference phases, including their data operations, i.e., data collection, transfer, storage and pre-processing. It is also required to measure the energy consumption/efficiency of ML entities at different model lifecycle phases and at data operation level for each of these phases. Energy consumption indicators should be introduced to be used to account for the total energy consumption of an ML entity. The indicator may be defined per phase or per more energy event, where one or more events may be associated to a model lifecycle phase. It is also important to keep track of total energy consumed in the entire AI/ML chain starting from AI/ML model training, storage and inference to training/inference data collection, transfer, processing, storage and access and total energy saved due to the decisions made based on the model output of each inference request. The ML MnS consumer may request the energy consumption/efficiency indicator from the ML MnS producer. Such energy consumption/efficiency indicators may be collected in a log, i.e. record of past events.

### 5.3.x.3 Potential requirements

**REQ-Energy\_MGT-01:** The AI/ML MnS Producer should have a capability allowing an authorized MnS consumer to provide its requirements on energy consumption/efficiency.

**REQ-Energy\_MGT-02:** The AI/ML MnS Producer should have a capability orchestrating a training (inference) solution including MLEntity(s) selection based on the energy consumption/efficiency requirements.

**REQ- Energy\_MGT-03: The** AI/ML MnS producer should have a capability to allow an authorized consumer to request and receive the evaluation and reporting of energy consumption/efficiency indicator values referring to all phases of an ML entity.

**REQ- Energy\_MGT-04:** The AI/ML MnS producer should enable an authorized MnS consumer to provide requirements on the energy consumption/efficiency for different ML lifecycle phases (e.g. training, inference) of one or more ML entities.

Note: The granularity for which energy consumption/efficiency requirements can be provided (e.g. per ML entity, ML inference function, ML training function, etc) is FFS

**REQ- Energy\_MGT-05:** The AI/ML MnS producer should enable MnS consumer to request for and receive the value of energy consumption/efficiency indicator based on the requirements.

**REQ- Energy\_MGT-06:** The AI/ML MnS producer should enable an authorized MnS consumer to request and receive.

***End of First change***