Contribution ID: **6** Type: **Demonstration/Tutorial (30 mins)**

AI Models as a Service with OSCAR: Integration with dCache and EGI Notebooks

Thursday, 22 June 2023 10:30 (20 minutes)

OSCAR is an open-source platform built on Kubernetes for serverless event-driven data-processing applications. To trigger the execution of these applications, OSCAR receives events from object-storage providers such as MinIO.

An OSCAR cluster can be easily deployed on the EGI Federated Cloud, as well as many other public and on-premises Clouds, through the Infrastructure Manager (IM), both integrated in the EOSC Portal.

Our past contribution demonstrated the capabilities of OSCAR, through the execution of different use cases, to create event-driven data-processing workflows along the computing continuum, processing part of the workload on the Edge and delegating the compute-intensive processing on a public cloud such as EGI Federated Cloud and Amazon Web Services (AWS).

In this contribution, we showcase, on the one hand, the integration of OSCAR with the distributed system for scientific data storage dCache, using it as an event source to trigger the execution of services. Using Apache Nifi to manage the event ingestion workflow between dCache and OSCAR, we have created a use case involving AI/ML models from the Deep Open Catalog for scalable asynchronous inference.

Moreover, we implemented a new Python API to trigger OSCAR services from Jupyter Notebooks, such as EGI Notebooks. We prepared another demo from EGI Notebooks to showcase interactive synchronous inference of AI/ML models.

Project PDC2021-120844-I00 funded by MCIN/AEI/10.13039/501100011033 funded by the European Union NextGenerationEU/PRTR. Also, grant PID2020-113126RB-I00 funded by MCIN/AEI/10.13039/501100011033. This work was supported by the project "interTwin"which has received funding from the European Union's Horizon Europe Programme under Grant 101058386. Also, by the project AI4EOSC "Artificial Intelligence for the European Open Science Cloud"which has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant 101058593

Other key topic

Machine Learning/Artificial Intelligence

Key Topic

Federated compute continuum

Primary authors: ALARCON MARIN, Caterina (Universitat Politècnica de València); MOLTO, German (Universitat Politècnica de València); CABALLER, Miguel (Universitat Politècnica de València); Mr LANGARITA BENITEZ, Sergio (Universitat Politècnica de València)

Presenters: ALARCON MARIN, Caterina (Universitat Politècnica de València); MOLTO, German (Universitat Politècnica de València)

Session Classification: Demonstrations