Contribution ID: **33** Type: **Demonstration**

Serverless workflows along the computing continuum with OSCAR/SCAR: Use cases from AI/ML inference

Wednesday, 21 September 2022 12:30 (25 minutes)

OSCAR is an open-source platform that supports serverless computing for event-driven data-processing applications. It abstracts away the deployment and management of computing resources through elastic Kubernetes clusters. Thanks to its integration with the Infrastructure Manager (IM), deployed as part of the European Open Science Cloud (EOSC), users can self-deploy these clusters on public and on-premises Clouds, including the EGI Federated Cloud.

It supports object-storage systems such as MinIO to trigger the execution of container-based services on file uploads, and the EGI DataHub, for mid-term data storage based on Onedata. Moreover, OSCAR can run on minified ARM-based clusters via K3s, thus making it possible to run it on the Edge.

Last year, we created new use-case examples and added new functionalities, such as the integration with Knative to improve auto-scaling in synchronous invocations, integration with Apache YuniKorn, support for private registries, and the ability to re-schedule jobs to service replicas.

OSCAR is integrated with SCAR, an open-source tool that pioneered the usage of containers within AWS Lambda. A common YAML-based Functions Definition Language (FDL) is available to define workflows, with a composer tool that simplifies its production. Latest releases of SCAR include support to mount EFS volumes and the usage of Amazon ECR to support larger container images.

In this contribution, we plan to demonstrate the benefits of the combination of OSCAR/SCAR to support event-driven data-processing workflows along the computing continuum, where partial processing can take place in the Edge and additional compute-intensive processing can take place on the EGI Federated Cloud and AWS. For this, several use cases will be demonstrated from the field of AI/ML, such as text-to-speech conversion, synchronous inference of ML models existing in the Deep Open Catalog or mask detection in public crowds, the latter exemplified as part of the AI-SPRINT project.

Any relevant links

OSCAR (web) - https://oscar.grycap.net
OSCAR (web - use cases) - https://oscar.grycap.net/blog/
OSCAR (GitHub repo) - https://github.com/grycap/oscar
FDL Composer - https://composer.oscar.grycap.net
Infrastructure Manager (Dashboard) - https://appsgrycap.i3m.upv.es:31443/im-dashboard/login
DEEP Open Catalog - https://marketplace.deep-hybrid-datacloud.eu
SCAR (web) - https://scar.readthedocs.io
SCAR (GitHub repo) - https://github.com/grycap/scar

Topic

A Federated Compute Continuum

Primary authors: RISCO, Sebastián (Universitat Politècnica de València); Dr MOLTO, German (Universitat Politècnica de València); Dr CABALLER, Miguel (Universitat Politècnica de València); ALARCÓN, Caterina (Universitat Politècnica de València); LANGARITA, Sergio (Universitat Politècnica de València)

Presenter: RISCO, Sebastián (Universitat Politècnica de València)

Session Classification: Demonstrations

Track Classification: A Federated Compute Continuum